REPUBLIQUE DU SÉNÉGAL
Un Peuple-Un But-Une Foi

MINISTÈRE DE L'ÉCONOMIE
ET DES FINANCES

MINISTÈRE DE L'AGRICULTURE
ET DE LA PISCICULTURE

ÉCOLE NATIONALE DE LA STATISTIQUE
ET DE L'ANALYSE ÉCONOMIQUE

INSTITUT SÉNÉGALAIS
DE RECHERCHES AGRICOLES

MÉMOIRE PROFESSIONNEL

PRÉDICTION DE LA DISTRIBUTION DES GLOSSINES DANS LA ZONE DES NIAYES PAR L'UTILISATION DE TÉLÉDÉTECTION

Présenté par :
DICKO Ahmadou
ÉLÈVE INGÉNIEUR
STATISTICIEN ECONOMISTE

Sous la direction de :
M. FALL AbdoulAziz
DOCTEUR, CIRAD
M. BOUYER Jérémy
DOCTEUR, HDR, CIRAD

Novembre 2011

Prédiction de la distribution des glossines dans la zone des Niayes par l'utilisation de télédetection

DICKO Ahmadou

Novembre 2011

Remerciements

Je témoigne toute ma gratitude à Jérémy Bouyer qui a co-encadré ce travail de stage, il a été très patient et pédagogue avec moi. J'ai appris avec lui une chose importante : *le travail d'équipe*, je le remercie pour tous ses conseils.

Je remercie Abdoul-Aziz Fall, co-encadreur de ce stage avec qui j'ai enrichi mes connaissances, il a contribué de manière significative aux résultats que j'ai obtenus sur ce document.

Je remercie Renaud Lancelot (CIRAD, UMR 15 CMAEE) pour avoir suivi de très près mes travaux, sa rigueur, son sens de la méthodologie et sa grande humilité en font un exemple. Avec Jérémy Bouyer, il fait partie de ceux qui m'ont donné envie d'embrasser la carrière d'enseignant chercheur.

Je remercie Jean Yves Rey (CIRAD, UPR 103 Hortsys) pour nous avoir fourni des données et pour m'avoir initié à mes premiers pas sur le terrain, j'ai essayé tant soit peu de profiter de sa très grande expérience.

Je tiens aussi à remercier l'équipe de zoologie de l'université d'Oxford, en particulier David Benz, pour nous avoir fourni des images satellites de très bonne qualité et je salue aussi la patience et la pédagogie dont il a fait preuve lors de nos échanges.

Ce travail n'aurait pas été possible sans l'effort conjugué de deux institut : l'Institut Sénégalais de Recheche Agricole (ISRA), et en particulier le Laboratoire National de d'Elevage (LNERV) et l'École Nationale de la Statistique et de l'Analyse Economique (ENSAE-Sénégal).

Nous remercions M. Macoumba DIOP directeur général de l'ISRA sans qui ce travail de stage n'aurait pas été possible.

Nous remercions M. THIONGANE Yaya pour nous avoir accepté au sein du LNERV et qui nous a permis de travailler dans un environnement scientifique de qualité.

Je remercie M. SECK Momar Talla chef du service de parasitologie du LNERV pour les différents échanges que nous avons eus et pour sa disponibilité.

Je remercie les membres du LNERV pour m'avoir mis dans des conditions exceptionnelles de travail durant mon stage.

Nous remercions M. TOURE Bocar directeur général de l'ENSAE Sénégal, pour sa disponibilité et les efforts qu'il a fourni pour nous assurer une formation de qualité.

Ce travail n'aurait pas été possible sans la contribution de notre responsable de filière M. CISSÉ Mamadou qui n'a ménagé aucun effort tout au long de notre formation pour nous mettre dans des conditions optimales de travail.

Je serai toujours redevable à l'administration de l'ENSAE-Sénégal pour son dévouement, l'ENSAE a changé à jamais ma façon d'aborder le travail, j'y ai appris la rigueur et le sérieux.

Je tiens aussi à remercier ceux qui m'ont aidé à la relecture du document final, en particulier Ba Khady qui n'a pas compté ses heures pour relire mes multiples fautes.

Je remercie aussi mon frère Ali Dicko pour les différents conseils qu'il me prodigue régulièrement, sur la façon d'aborder la recherche.

Je tiens à dire merci à ma soeur Fatma Dicko qui est pour moi le plus grand exemple de courage et de ténacité.

Je remercie enfin mon père, Hamady Y. Dicko qui a toujours cru en moi et dont le soutien inconditionnel me permet de donner chaque jour le meilleur de moi.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ACP	Analyse en Composante Principale
AVHRR	Advanced Very High Resolution Radiometer
CIRAD	Centre International de Recherche en Agronomie pour le Développement
ENFA	Ecological Niche Factor Analysis
ENSAE	École Nationale de la Statistique et de l'Analyse Économique
FAO	Food and Agriculture Organization
ISRA	Institut Sénégal de Recherche Agricole
LST	Land Surface temperature
MADIFA	Mahalanobis Distance Factor Analysis
MaxEnt	Maximum entropy
MIR	Middle Infrared
MODIS	Moderate resolution Imaging Spectroradiometer
NDVI	Normalized Differenced Vegetation Index
NOAA	National Oceanic and Atmospheric Administration
PAAT-IS	Program against African Trypanosomiasis - Information System
PATTEC	Pan African Tsetse and Tryponosomiasis Eradication Campaign
SIG	Système d'Information Géographique
TAA	Trypanosomes Animale Africaine

Abstract

African animal trypanosomiasis (AAT) are a major pathological constraints to livestock in the Niayes (Senegal). In the Niayes area, the main vector of this disease, *Glossina palpalis gambiensis* (Diptera : Glossinidae) is present and this species is subject to an eradication campaign launched in 2005. The aim of the study is to establish a spatial distribution of the species in order to define an infested area. This infested area will be used in a targeted control. An approach based on ecological knowledge of the species has been developed to fight against this vector borne-disease. We used several species distribution models to understand the impact of environmental variable on the species. But we also used those models to map the potential distribution of *G. p. gambiensis* through our area of study. Theoretical knowledge on the species were operated through the theory of ecological niches from which distribution models are based. The data required for modeling were obtained using satellite image (MODIS). This time series of images were used to derive climatic and environmental data which could be linked to tsetse fly distribution. The analysis of the species responses to these variables have been made through exploratory analysis based on the concept of ecological niche that allowed to describe and select variables that influence the most the distribution of the vector of the AAT. The Results showed a high sensitivity to variables related to vegetation. These variables were then used by three predictive models to delineate the targeted area. Among those models, two have a good predictive power (AUC = 0.8), and the best of them (MaxEnt) were used to obtain results that are consistent with the ecology of the species. Hopefully, the former control area are not going to change a lot. However, these models showed some suitable area where there were no sampling.

Keywords : *species distribution models, remote sensing, ecological niche, riparian glossina*

Résumé

Les trypanosomoses animales africaines (TAA) sont une des principales contraintes pathologiques à l'élevage dans la zone des Niayes (Sénégal). Dans la zone des Niayes, le seul vecteur cyclique de cette maladie, *Glossina palpalis gambiensis* (Diptera :Glossinidae) est présent, il est l'objet d'une campagne d'éradication lancée depuis 2005. Cette étude a pour objectif d'établir une distribution de cette espèce, dans le but de définir une zone infestée qui servira à une lutte ciblée. Une approche basée sur la connaissance écologique de l'espèce à été développée pour lutter contre cette maladie vectorielle. Nous avons utilisé plusieurs modèles de distribution d'espèces afin de comprendre l'impact des variables environnementales sur l'espèce d'une part et d'autre part d'établir une cartographie de sa distribution spatiale potentielle à l'échelle de notre zone d'étude. Les connaissances théoriques sur l'espèce, ont été exploitées par le biais de la théorie des niches écologiques sur laquelle se basent les modèles de distribution d'espèce. Les données nécessaires à la modélisation ont été obtenues à l'aide d'images satellitales (MODIS). Il s'agit de séries temporelles d'images qui ont permis de mesurer à l'échelle de la zone, les phénomènes climatiques et environnementaux auxquels les glossines riveraines sont fortement sensibles. Pour analyser cette sensibilité, nous avons utilisé des analyses exploratoires basées sur le concept de niche écologique qui ont permis de décrire et choisir les variables qui influencent la distribution du vecteur de la TAA. Les résultats ont montré une forte sensibilité aux variables liées à la végétation. Ces variables ont ensuite été utilisées par trois modèles prédictifs afin de délimiter la zone de lutte. Parmi ces modèles, deux ont un bon pouvoir prédictif (AUC = 0.8), et le meilleur d'entre eux (MaxEnt) permet d'avoir des résultats concordant avec l'écologie de l'espèce. Globalement, la zone de lutte définie avant cette analyse n'est pas remise en cause. Cependant, ces modèles ont permis de cartographier des zones des Niayes non échantillonnées, où de toute vraisemblance *G. p. gambiensis* pourrait s'établir.

Mots-clés : *modèles de distribution d'espèce, télédétection, niche écologique, glossines riveraines*

Introduction générale

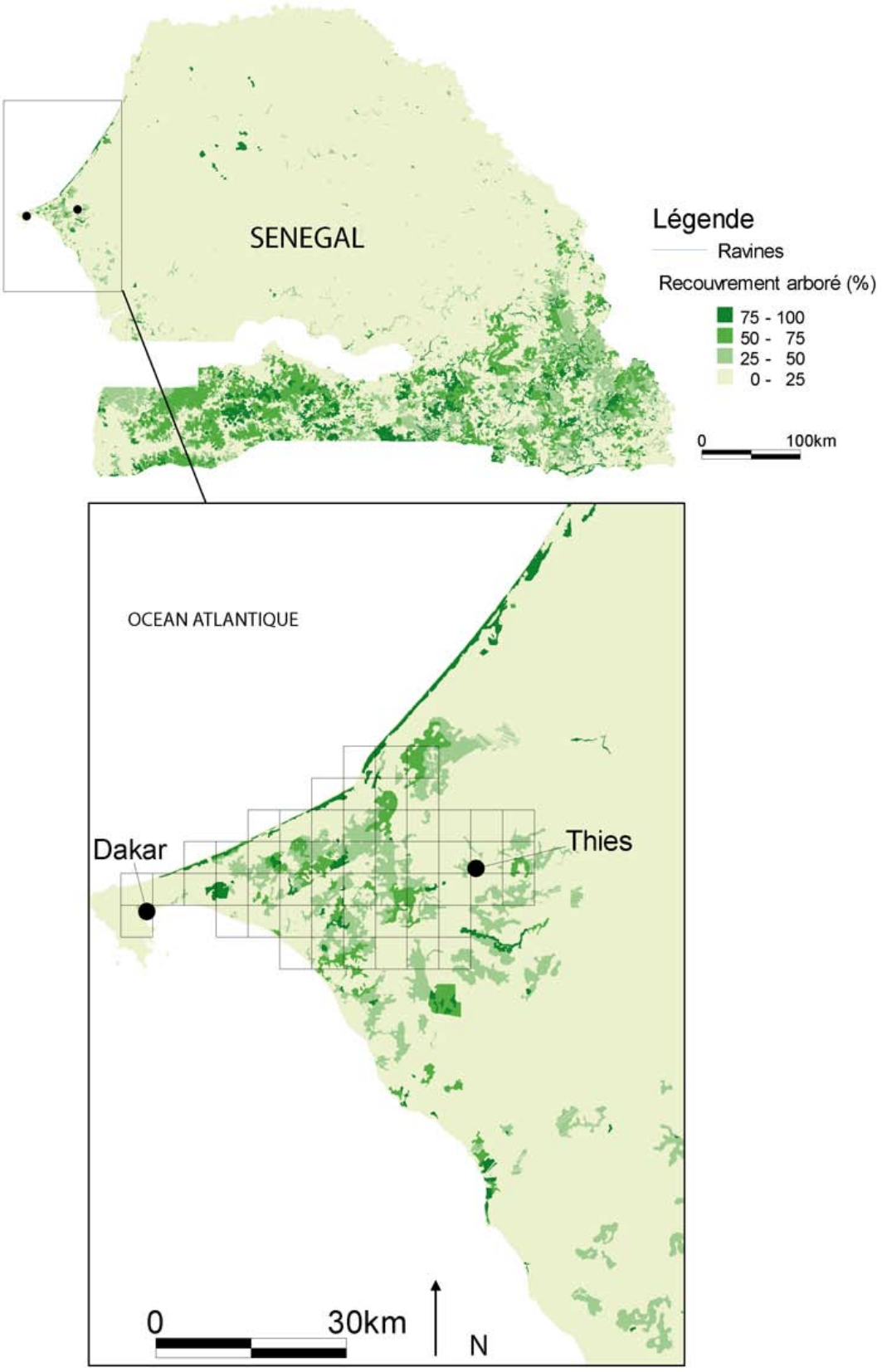
Les maladies à transmission vectorielle sont des maladies pour lesquelles l'agent pathogène est transmis d'un individu (en général vertébré) à un autre par l'intermédiaire d'un arthropode. Ces arthropodes sont sensibles aux variations des conditions climatiques comme la température ou encore l'humidité. Chacun de ces facteurs climatiques affecte directement la mortalité, les flux migratoires et la fécondité de l'espèce. Ils jouent donc un rôle majeur sur la taille de la population et sa localisation.

Le contrôle des maladies vectorielles constitue aujourd'hui un enjeu majeur de développement. Une part importante de ce contrôle repose sur le contrôle de leurs vecteurs, en particulier dans le cas des trypanosomoses africaines, objets de la présente étude. Le contrôle des vecteurs passe par une compréhension de l'écologie de l'espèce ciblée afin de mettre en place des méthodes adaptées en optimisant le rapport coût-bénéfice des campagnes de lutte. Cela est encore plus vrai dans le cas d'une campagne d'éradication des vecteurs, où toutes les sous-populations doivent être touchées. En effet, oublier une poche infestée peut conduire à la ré-infestation complète des zones libérées, annulant ainsi la rentabilité de la campagne.

Au Sénégal, le gouvernement a lancé depuis 2005 une campagne d'éradication des glossines dans la zone des Niayes, à forte potentialité de production agricole et animale. Cette campagne a commencé par une étude de faisabilité incluant une étude entomologique large pour cibler la zone de lutte. En tous, 683 sites ont été prospectés permettant d'obtenir une délimitation de la zone de lutte (Bouyer *et al.*, 2010b). Cependant, ce type d'enquête est très coûteux et le nombre de sites enquêtés est donc limité. Les modèles de distribution d'espèce sont des outils de choix pour l'aide à la décision dans cette situation car ils permettent d'espérer une meilleure définition de la zone infestée à partir de ces données de présence/absence.

L'objet de notre travail est donc d'obtenir une estimation plus fine et objective de la zone infestée par les glossines que celles réalisée à partir de l'enquête entomologique préalable.

Choix de la zone d'étude



Graphique 0.1 : Zones des Niayes, Sénégal. Les grilles représentent la zone de lutte a priori (avant modélisation).

Au Sénégal, la zone côtière des Niayes (graphique 0.1) présente un fort potentiel laitier grâce à des conditions environnementales et économiques favorables. L'élevage bovin laitier s'y intensifie, avec des fermes produisant 3000 à 5000 litres par jour. Les éleveurs, souvent sédentaires, bénéficient d'un programme d'amélioration génétique, avec plus de 32 000 inséminations artificielles réalisées à ce jour. Cependant, les animaux des petits producteurs sont souvent croisés avec des taurins Ndama trypanotolérants [1] mais peu productifs (Seck *et al.*, 2010). En effet, les conditions climatiques ont permis la persistance d'un vecteur des trypanosomoses animales, *Glossina palpalis gambiensis* (Diptera, Glossinidae) et une forte incidence de la trypanosomose bovine. Suite à une étude de faisabilité (2007-2010), la zone infestée a été délimitée à environ 1 000 km^2 et l'isolement complet de la population cible a été confirmée par génétique des populations. Une stratégie d'éradication a été décidée et mise en place par les services vétérinaires en 2011 (Bouyer *et al.*, 2010b). Après une phase de réduction de la densité des glossines par traitement épicutané des bovins et la pose d'écrans imprégnés d'insecticide, l'éradication devrait être obtenue par lâchers aériens de glossines mâles stériles (Bouyer *et al.*, 2010c).

Épidémiologie des trypanosomoses animales africaines

Les trypanosomoses sont des maladies parasitaires, à transmission vectorielle, déclenchées par les trypanosomes dans un organisme hôte. Le trypanosome est un parasite pathogène transmis en Afrique par les glossines et en Amérique latine par des punaises. Les trypanosomoses africaines sont transmises à la fois aux hommes et au bétail. La trypanosomose humaine africaine (THA), encore appelée « maladie du sommeil » sévit exclusivement en Afrique subsaharienne. Selon l'OMS en 2000, plus de 55 millions de personnes étaient soumises au risque de contracter la maladie dans les 37 pays qui courent le risque de THA. Depuis 1995, on enregistre une recrudescence telle qu'on évalue à 500000 le nombre de malades supplémentaires par an, dont 95% ne sont ni suivis, ni traités, entraînant le décès de 100 personnes par jour. Cependant, ces dernières années, suite à des campagnes d'éradication et à l'augmentation des prospections médicales le nombre de cas a diminué. Selon l'OMS, en 2006 le nombre de cas de THA a été estimé à 70000. L'objet de notre étude, n'est pas la THA, mais la trypanosomose animale africaine (TAA) qui touche principalement le bétail domestique dans les zones agropastorales.

Les TAA sont des maladies infectieuses, inoculables, non contagieuses, qui évoluent le plus souvent sous forme chronique anémiante conduisant à la cachexie [2] et à la mort. L'organisation des Nations-Unies pour l'alimentation et l'agriculture (FAO) estimait, en 1997, que les pertes économiques dues aux trypanosomoses animales étaient de l'ordre de 1 à 1.5

1. capacité du bétail à produire, survivre à la TAA sans l'aide de chimiothérapie
2. dégradation profonde de l'état général, accompagnée d'une maigreur importante

milliards d'Euro par an. Globalement, les trypanosomoses animales réduiraient le nombre de tête de bétail de 10 à 50 % et la production agricole de 2 à 10 % dans les zones infestées par les tsé-tsé (Itard *et al.*, 2003).

Le vecteur : la glossine

En Afrique, la transmission du trypanosome à l'hôte mammifère est réalisée par un insecte diptère hématophage, la glossine ou mouche tsé-tsé. Cette mouche d'environ de 1 cm de long est utilisée par un parasite de la taille d'un globule rouge.
Les glossines sont des insectes Diptères cyclorraphes appartenant à la famille des Glossinidae et au genre glossina. On compte trois sous-genres et 31 espèces et sous-espèces de glossines, elles sont toutes susceptibles de transmettre des trypanosomes. L'espèce étudiée dans ce document, *G. palpalis gambiensis* fait partie du groupe *palpalis*. C'est la seule espèce présente dans les Niayes.

Graphique 0.2 : *G. palpalis gambiensis*. crédit photo : Bouyer J.

G. p. gambiensis (graphique 0.2) est un vecteur majeur de la THA et de la TAA en Afrique de l'Ouest. Elle possède une grande capacité d'adaptation et à la fragmentation et à a la dégradation de son habitat d'origine : les forêt galéries où les savanes de zones éco-climatiques soudaniennes ou guinéennes, respectivement.

Plan

Objectifs de recherche

Les objectifs principaux de ce travail de stage est d'obtenir la distribution géographique potentielle de *G. p. gambiensis* dans la zone des Niayes à l'aide de variables environnementales obtenues principalement par télédétection. Plus spécifiquement, il s'agit de :

1. Établir une carte de distribution potentielle en utilisant plusieurs approches.

2. Comparer les différents modèles et le pouvoir prédictif des cartes réalisées.
3. Identifier les variables environnementales les plus importantes pour l'espèce.
4. Vérifier si les résultats sont en concordance avec les connaissances théoriques sur l'espèce.

Questions de recherche

1. Est-il possible d'obtenir une cartographie des zones favorables sous formes de distribution de probabilité dans la zone des Niayes ?
2. Existe-t-il une différence significative entre les différentes cartes et comment mesurer cette différence ?
3. Quels sont les variables environnementales les plus importantes dans le choix de l'habitat de *G. p. gambiensis* ?
4. Parmi les modèles utilisés lesquels sont complémentaires ?
5. Les conclusions obtenues sont elles en accord avec l'écologie de l'espèce ?
6. La zone de lutte initiale doit-elle être remise en cause ?

Plan

Ce document sera divisé en quatre principaux chapitres :

1. Le premier chapitre présente quelques bases théoriques qui seront utilisé par la suite, il s'agira d'une approche bibliographique présentant dans une première partie la théorie écologique qui sous-tend la modélisation de la distribution des espèces et dans une seconde partie on présentera les différents types de modèles rencontrés dans la littérature ;
2. Dans le second chapitre, les données qui seront utilisées dans les différentes analyses sont présentées, ainsi qu'une explication des modifications qu'elles ont subies ;
3. Le troisième chapitre introduit les premières analyses, il s'agira principalement de présenter de nouvelles méthodes d'analyse exploratoire des données adaptées à la modélisation de la distribution d'espèce ;
4. Enfin, le dernier chapitre sera consacré à la mise en place de plusieurs modèles prédictifs de l'habitat de l'espèce étudiée. Ces modèle tireront profit de l'analyse exploratoire antérieure.

Chapitre 1

Niches écologiques et modèles de distribution d'espèce

Dans ce chapitre introductif, on définit les bases théoriques nécessaires à l'étude de la niche écologique d'une espèce, et de sa distribution spatiale. Le succès de la modélisation dépend en grande partie de la conceptualisation et de la traduction des phénomènes écologiques sous une forme adaptée à la mise en place de modèles. En écologie quantitative, la théorie des niches écologiques permet de répondre à cette préoccupation.
Dans ce chapitre, on définira la notion de niche écologique et son importance dans les modèles de distribution des espèces. On mettra aussi en relief les particularités de l'écologie de *G. p. gambiensis*, vecteur des trypanosomes causant les TAA au Sénégal.
Dans une seconde partie, on expliquera la particularité des modèles de distribution d'espèce. Cette description sera suivie d'une revue de quelques études menées par d'autres auteurs sur la modélisation de la distribution des glossines.

1.1 Théorie des niches écologiques

1.1.1 Qu'est ce qu'une niche ?

Le concept de niche écologique est central en écologie et remonte aux travaux de Grinnell (1917) qui l'a définie alors comme les facteurs environnementaux minimum dont a besoin une espèce pour survivre sans immigrer. En effet, en l'absence d'immigration, une espèce donnée ne peut survivre que si la combinaison de variables environnementales dans la zone qui l'entoure permet une croissance de sa population (Hirzel et Le Lay, 2008) . Le

concept de niche Grinnellienne contraste avec celui de niche Eltonienne (Elton, 1927 ; Soberón, 2007), en ce sens que Elton (1927) définit une niche comme la fonction d'une espèce dans son environnement biotique.
Hutchinson (1957) développe et affine ce concept autour de deux notions, la *niche fondamentale* et la *niche réalisée*. Il définit la première comme un *hypervolume* dont chaque dimension représente un des états de l'environnement qui permettrait à l'espèce d'exister (survivre et se reproduire), et la niche réalisée ne serait qu'un sous espace de la niche fondamentale résultant de l'interaction de l'espèce avec d'autres espèces (i.e compétition, prédation, mutualisme, etc.). La différence entre la niche fondamentale et la niche réalisée est très importante dans la mise en place du modèle conceptuel. Pour Whittaker *et al.* (1973), seule la niche réalisée peut être mesurée car il n'est pas possible d'exclure toutes les interactions liées aux facteurs biotiques lors d'expériences de terrain et de collecte d'information. Néanmoins ils soutiennent que l'aggrégation de toutes les niches réalisées devrait conduire à la niche fondamentale.
La notion de niche (*sensu* Hutchinson) est fondamentale car elle permet de faire le lien entres les connaissances dont on dispose sur l'espèce (démographie, comportement trophique, etc) et un modèle conceptuel géométrique, qui permet alors d'explorer et de modéliser cette niche en utilisant des outils quantitatifs. Cependant, avant la phase de modélisation, certaines hypothèses doivent être faites sur le système biologique étudié. Parmi ces conditions préalables, une des plus importantes est celle d'*équilibre*.

1.1.2 La notion d'équilibre

Une espèce est dite à l'équilibre avec son environnement si elle est présente dans toutes les zones favorables et absente des zones défavorables. Ce degré d'équilibre est étroitement lié à l'interaction avec les facteurs biotiques et la capacité de dispersion de l'espèce. Pour Guisan et Zimmermann (2000), un modèle basé sur l'hypothèse de non équilibre doit être dynamique et stochastique car il est difficile de modéliser la réponse d'une espèce à un environnement sujet à des changements dynamiques et stochastiques.
Cependant, on note que *G. p. gambiensis* est une espèce persitente bien établie dans la zone d'étude à cause de son caractère k-stratégiste [1]. En effet, les glossines font parties des rares insectes à adopter ce type de stratégie démographique, qui est habituellement observé chez les grands mammifères.
Finalement, dans le cadre de notre étude, l'hypothèse d'équilibre se révèle alors beaucoup moins restrictive et permet d'utiliser des modèles statistiques d'une grande complexité. Nous partirons dès lors du principe que *G. p. gambiensis* est une espèce en équilibre et la

1. une espèce adopte une stratégie de type K si elle est caractérisée par une faible fécondité, une faible mortalité, une faible densité et qu'elle est adaptée aux ressources disponibles dans son milieu (importance des facteurs de régulations densité dépendants)

modélisation de la sa niche par des modèles non dynamiques est alors possible.

Dans la littérature sur les théories écologiques, derrière les différents modèles de distribution d'espèce il existe plusieurs écoles et autant de nuances autour des concepts théoriques de base. On note qu'il existe un consensus autour de l'importance de la niche écologique comme base conceptuelle derrière les différents modèles. Cependant, certains auteurs préfèrent parler plutôt d'« habitat » quand ils s'exercent au même type de modélisation.

1.1.3 Niche et habitat : une question d'échelle

Les concepts de niche ou d'habitat sont souvent confondus et il n'existe aucun consensus sur les véritables différences entre ces deux notions. Cependant certains auteurs ont tenté de donner une définition plus ou moins complexe de ce qu'ils entendent par habitat. C'est ainsi que Morrison *et al.* (2006) définissent simplement un habitat comme étant « un endroit où vit l'espèce ». Pour Hall *et al.* (1997), l'habitat est défini par les ressources et les conditions présentes dans une zone qui permettent l'occupation (survie et reproduction inclus) de cette zone par une espèce. La définition que nous retiendrons et qui se rapproche le plus de la méthodologie que nous proposons est celle de Whittaker *et al.* (1973) pour qui la notion d'habitat est lié à celle de variables d'habitats qui forment selon lui un hyper-espace défini par un certain nombre de facteurs abiotiques, la partie de cet hyper espace, occupé par une espèce, étant appelé hypervolume d'habitat. Pour eux, la réponse d'une espèce aux variables d'habitat dans cet hypervolume, mesurée au niveau de sa population, décrit l' « habitat » de cette espèce.
Cependant, il faut noter dans cette définition que la notion d'hypervolume apparaît et rappelle alors l'analyse de Hutchinson (1957) sur la niche écologique. Néanmoins, les concepts de niche (*sensu* Hutchinson (1957)) et d'habitat ne sont pas à confondre. Afin d'éviter toute confusion avec la notion de niche, Whittaker *et al.* (1973) précisent que toutes les variables environnementales peuvent être séparées en deux groupes, celles d'habitat et celles de niche. Les variables d'habitat sont celles qui ont une forte composante spatiale (avec une échelle spatiale grossière) et qui sont en général définies à l'échelle de la communauté à laquelle appartient l'espèce. Les variables de niches, selon eux, décrivent le rôle fonctionnel d'une espèce au sein d'une communauté. Ces variables sont locales, généralement non spatialisées et sont définies à une échelle très fine. Ce dualisme introduit par Whittaker *et al.* (1973) correspond à celui introduit par Hutchinson (1978) qui distingue aussi deux types de variables pour décrire une niche : les variables *scénopoetiques* (habitats) et les variables *bionomiques* (niches). Il semble alors évident que cette nuance entre habitat et niche nous ramène aux notions de *macrohabitat* et de *microhabitat*, le premier étant en général défini à une échelle spatiale beaucoup moins fine que le second. Par analogie avec l'approche de

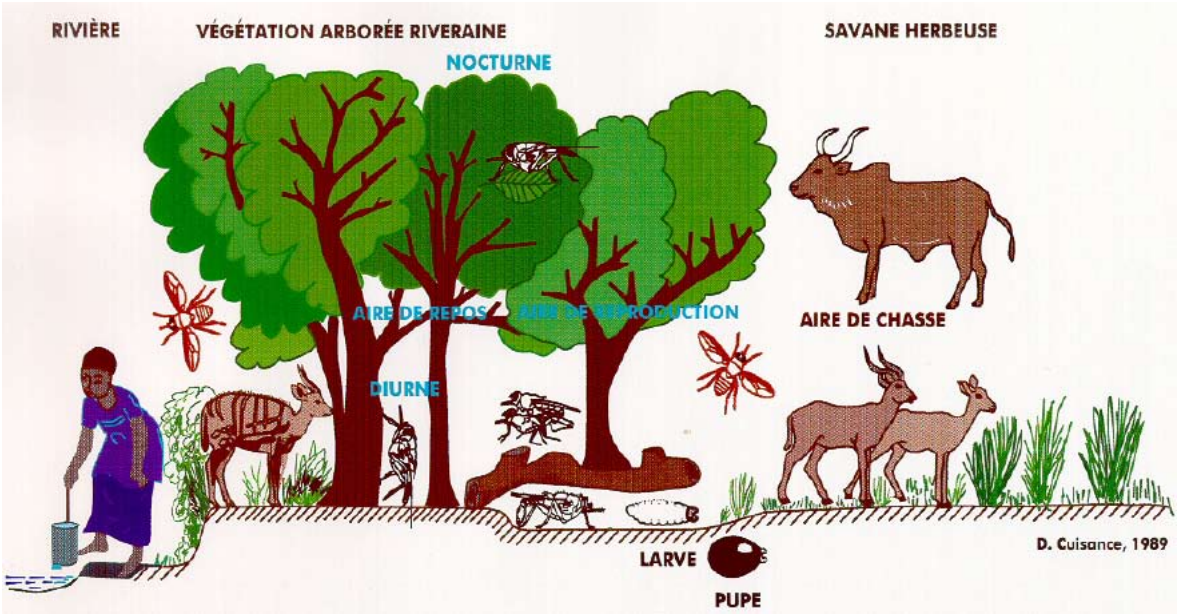
Hutchinson (1957), la notion d'habitat est alors proche de celle de niche réalisée et celle de niche (*sensu* Whittaker *et al.* (1973)) est plus proche de celle de la niche fondamentale. On peut aussi rajouter que la niche (*sensu* Whittaker *et al.* (1973)) rejoint la vision de Elton (1927) et la niche Grinnellienne décrit ce que Whittaker *et al.* (1973) appellent habitat.

Par la suite, nous désignerons par niche, la niche réalisée de l'espèce et donc d'une certaine manière son habitat.

1.2 Notions sur l'écologie des glossines

La glossine est une espèce très sensible aux conditions climatiques de son environnement (Nash, 1937, 1948). L'aire de répartition des glossines est marquée par une température annuelle moyenne supérieure à 20 °C et une pluviométrie supérieure à 400 mm. Les variations climatiques annuelles déclenchent des stratégies chez les glossines qui cherchent leur optimum écologique ; elles modifient l'amplitude de leur niche selon les conditions saisonnières. La dispersion des glossines riveraines [2] est importante en saison des pluies tandis qu'elles se concentrent dans les galeries forestières et les patches de végétation forestières humides durant les mois les plus chauds et secs. L'optimum hygrométrique varie de 70% à 81% pour G. p. *gambiensis* et son optimum thermique se situe entre 23 et 26 °C (Rogers, 1979 ; Challier et de Paris VI, 1973). Dans la zone des Niayes, Touré (1974) donne une description précise de l'habitat de *G. p. gambiensis*. Cette zone a la particularité d'être à l'extrême limite de l'aire de distribution de cette espèce au Sénégal, qui est en outre le pays le plus septentrional dans sa répartition géographique. Il note que l'habitat caractéristique de cette espèce est constitué de forêt humide et de cordons ripicoles. Cependant, il décrit aussi la présence de gîtes secondaires généralement caractérisés par une végétation buissonneuse assez dense, constituée principalement de vergers (agrumes, manguiers, etc.) et de haies d'euphorbes. Bouyer *et al.* (2010b) ont montré que ce second type d'habitats est devenu prédominant dans la zone d'étude.

2. espèce de glossine de galerie riveraine dont *G. p. gambiensis* fait parti



Graphique 1.1 : Ambit de la glossine (*sensu* Jackson)

Chez les glossines, les écologistes du comportement ont défini un type d'habitat particulier : *l'ambit* (Jackson, 1941) (graphique 1.1), qui correspond à un espace structuré en sites de reproduction et d'alimentation où la glossine se déplace grâce à sa mémoire topographique (Bouyer, 2006).
D'un point de vue démographique, les glossines sont des espèces à forte composante K, à ce titre elles présentent un cycle de développement long. Elles vivent en moyenne trois ou quatre mois avec un taux de reproduction très bas (au maximum, 10 descendants par femelle). De plus chaque femelle s'accouple une seule fois dans sa vie, et ce, même si le mâle est stérile.
Ce type de comportement démographique particulier, a des conséquences très importantes dans les campagnes de contrôle et de lutte. En effet, une mortalité journalière d'environ 3% des femelles résulterait en un déclin de la population des glossines sous contrôle (Hargrove, 1988). De plus, le comportement sexuel et la faible fécondité des femelles rendent très efficaces les techniques de lâchers de mâle stérile.

L'écologie des glossines riveraines est très complexe et pour une vision d'ensemble plus détaillée, Bouyer (2006) donne une revue assez récente. La sensibilité des glossines aux variations des conditions climatiques (Nash, 1937, 1948), leurs comportements démographiques singuliers sont autant d'informations qui peuvent permettre de modéliser leurs distributions spatiales car elles nous permettent d'avoir une idée des variables environnementales qui peuvent influencer sa niche. Les connaissances dont on dispose sur les

glossines riveraines et sur *G. p. gambiensis* nous permettent alors de comprendre les variables environnementales et climatiques qui influencent sa niche. Ces connaissances et la théorie de la niche écologique rendent alors possible la mise en place de modèles analytiques. Ces modèles complexes permettront alors d'explorer et d'analyser la niche de *G. p. gambiensis*.

1.3 Modèle de distribution d'espèce

La modélisation de la distribution spatiale des espèces est une discipline récente qui doit son émergence au développement des systèmes d'informations géographiques (SIG) et des méthodes statistiques. Ils s'agit de modèles basés sur le concept de niche écologique (*sensu* Hutchinson (1957)). Le but est d'identifier les conditions optimales qui permettent à une espèce de se maintenir et/ou d'étudier l'impact écologique de certains facteurs (anthropisation, changement climatique, etc.) sur l'espèce. On distingue de manière classique deux approches : l'approche mécaniste et l'approche corrélative.

Les modèles mécanistes sont basés sur une approche mathématique (représentation de la dynamique par des équations différentielles résolues en temps continu ou discret) ou informatique (intéraction entre acteurs du système et avec l'environnement définies par des règles sémantiques traduites en algorithmes puis en programme informatiques). Ces modèles utilisent la réponse que l'espèce donne à des variables environnementales en tenant compte de certaines variables démographiques liées à l'espèce (fécondité, mortalité, etc.) et de leurs transitions ou dynamiques. Ce type de modèle demande un très bonne connaissance de l'espèce étudiée (traits de vie et écologie).

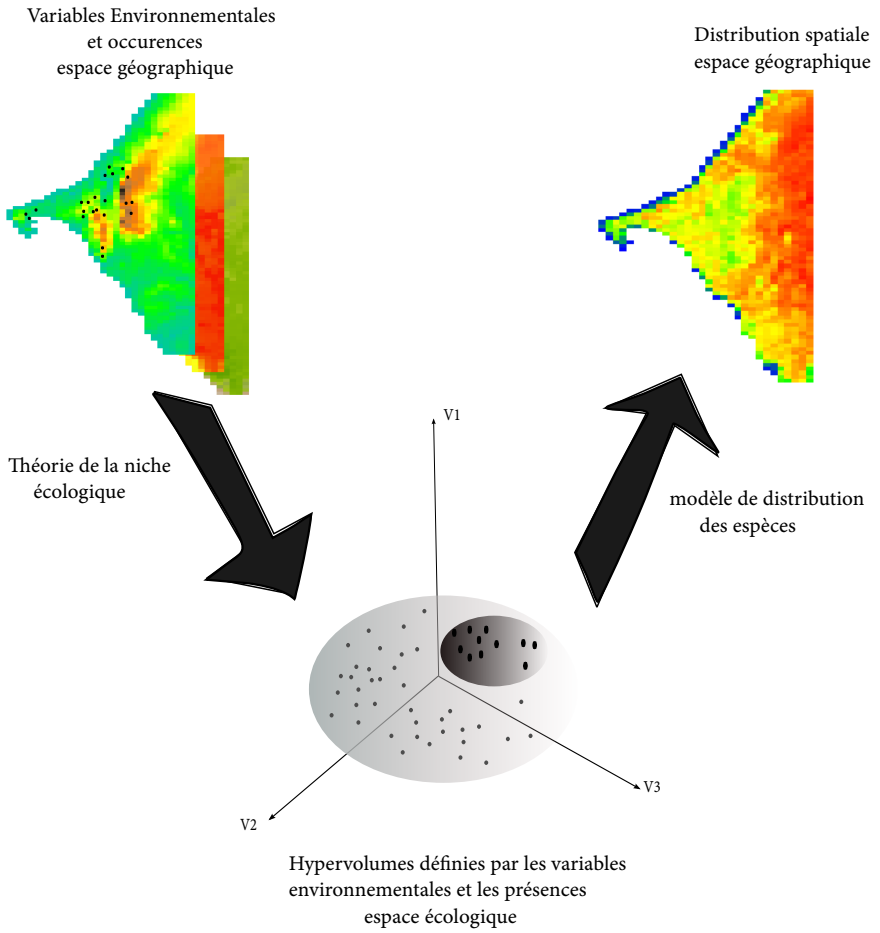
Les modèles corrélatifs quant à eux, estiment les conditions optimales à l'établissement d'une espèce en associant les données de présence (ou/et d'absence) aux variables environnementales qui doivent être alors bien choisies afin d'avoir une influence sur la niche de l'espèce. Les modèles corrélatifs permettent ainsi d'obtenir une prédiction spatiale des zones favorables au maintien de l'espèce étudiée. D'un point de vue analytique, ce type de modèle tire leurs bases théoriques des méthodes statistiques sous-jacentes (classifications supervisées, régressions, etc.). Les modèles que nous utiliserons dans la suite de ce document font partie de cette famille de modèles.

Les modèles de distribution d'espèce sont des outils importants d'aide à la décision. Ils ne sont pas seulement utilisés dans le cadre de campagne d'éradication. C'est dans le domaine de la conservation des espèces que cette discipline a émergé. Parmi, les utilisations qu'on peut faire de ce type de modèle on peut citer :

- La conservation d'espèces en danger ;
- la découverte de nouvelles espèces ;

– l'impact des variables environnementales et leurs variations sur une ou plusieurs espèces ;
– la lutte intégrée et les projets d'éradications.

Néanmoins la principale information qu'on obtient avec ce type de modèle est la relation qui existe entre une espèce et les différents facteurs biotiques et abiotiques qui composent son environnement (Austin, 2002, 2007). La modélisation de cette interaction entre une espèce et son environnement permet alors d'obtenir des cartes de prédictions des zones favorables au maintien de l'espèce par exemple.
Il existe une multitude de modèles différents qui sont souvent comparés entre eux (Elith *et al.*, 2006). Le but final reste le même mais les approches sont souvent différentes et dépendent beaucoup de la disponibilité des données, l'adaptation des algorithmes aux cadres des données écologiques.



Graphique 1.2 : Relation entre l'espace géographique et la niche écologique, on passe de l'espace géographique qui contient les variables explicatives à l'espace écologique à travers le concept théorique de niche. Cet espace est propice à l'utilisation de modèle complexe qui permettent d'obtenir une cartographie de la niche de l'espèce.

Afin de mieux comprendre ces modèles, il convient de distinguer deux espaces : l'*espace écologique* et l'*espace géographique*. L'espace écologique est lié a la notion de niche et dépend du choix des variables environnementales qui la caractérisent. L'espace géographique est une projection de l'espace écologique dans une région particulière. Cet espace est constitué de cellules ou pixels qui couvrent cette région. De manière classique, on procède en modélisant dans un premier temps la niche dans l'espace écologique, ensuite on projette le résultat final sur l'espace géographique (graphique 1.2).

Parmi les différences qui existent entre les différents modèles, une des plus importantes est le type de données de terrain qu'on utilise pour construire ces modèles. Cette distinction mène à une classification de ces algorithmes. En général, nous distinguons :

1. Les modèles de présence-seule
2. Les modèles de présence-absence
3. Les modèles de présence-background
4. Les modèles de présence-pseudoabsence

1.3.1 Modèles de présence-seule

Ce sont des approches basées uniquement sur l'utilisation des données de présence sans autres considérations sur l'espace disponible. Il existe deux approches : une approche basée sur la construction d'enveloppes et une seconde basée sur l'utilisation de distances dans l'espace écologique. Les modèles d'enveloppes sont des méthodes simples pour estimer la niche de l'espèce, on peut citer par exemple l'algorithme HABITAT qui construit une enveloppe convexe autour des données de présence. Les méthodes basées sur les distances quant à eux sont un peu plus complexes en général. Le principe est de calculer un indice de dissimilarité entre des points dans l'espace écologique. La Distance de Mahalanobis (chapitre 4) est très souvent utilisée dans ce cas précis, elle mesure la distance entre les données de présence et l'optimum de la niche. Cet optimum est souvent représenté par la moyenne des variables sur l'espace occupé par l'espèce (niche).

1.3.2 Modèles de présence-absence

Quand les données sur les *vraies absences* (section 2.3.2) sont disponibles, l'estimation de la distribution géographique d'une espèce est possible grâce à l'utilisation de méthodes statistiques plus conventionnelles qui permettent de discriminer les présences des absences. Ces approches sont souvent basées sur des modèles de régressions comme les modèles linéaires généralisés (GLM), les modèles additifs généralisés (GAM), etc. Une autre approche consiste à utiliser des algorithmes d'apprentissages statistiques comme les arbres

de décision, les réseaux de neurones ou encore les méthodes d'ensemble comme les forêts aléatoires (chapitre 4).
Pour Brotons *et al.* (2004), l'information que recèlent les données d'absence est très importante et devrait être incluse dans les modèles suivant leur disponibilité.

1.3.3 Modèles de présence-background

On parle de données de « background » pour décrire l'environnement disponible sur lequel se fait l'analyse. Les méthodes qui utilisent cette information dans leurs analyses sont différents des celles de présence-absence et de présence-seule, en ce sens qu'elles comparent les zones utilisées par l'espèce à celles disponibles afin de modéliser les conditions favorables pour l'espèce. Une des méthodes de présence-background la plus utilisée est MaxEnt (chapitre 4). MaxEnt est basée sur un principe fondamental de le théorie de l'information, celui de l'entropie maximum. D'autres méthodes factorielles comme l'Analyse Factorielle des Niches Ecologiques (ENFA) (chapitre 3) ou encore la décomposition Factorielle des Distances de Mahalanobis (MADIFA) (chapitre 3) appartiennent aussi à ce type d'algorithme. Ces modèles factoriels cherchent des directions (plan) dans lesquelles la projection de la niche serait optimale. Il s'agit de méthodes adaptées à l'analyse exploratoire de la niche d'une espèce (Calenge et Basille, 2008).

1.3.4 Modèles de présence-pseudoabsence

Quand on ne dispose pas de données de présence, une approche consiste à considérer des points de l'espace disponible où ne se trouvent pas l'espèce comme des *absences*. Cette hypothèse, permet alors d'utiliser des méthodes de discrimination classique. Le choix de ces pseudo-absences est important pour la réussite des différents modèles utilisés (Chefaoui et Lobo, 2008).

Il faut cependant noter que quelle que soit le type de méthode utilisée, des précautions sont à prendre pendant la phase de modélisation.

1.3.5 Quelques considérations importantes

Ces modèles sont sensibles à plusieurs facteurs. Il faut donc prendre certaines considérations lors de leurs utilisations. Selon Soberón et Peterson (2005) il existe quatre facteurs qui influencent la distribution spatiale d'une espèce :

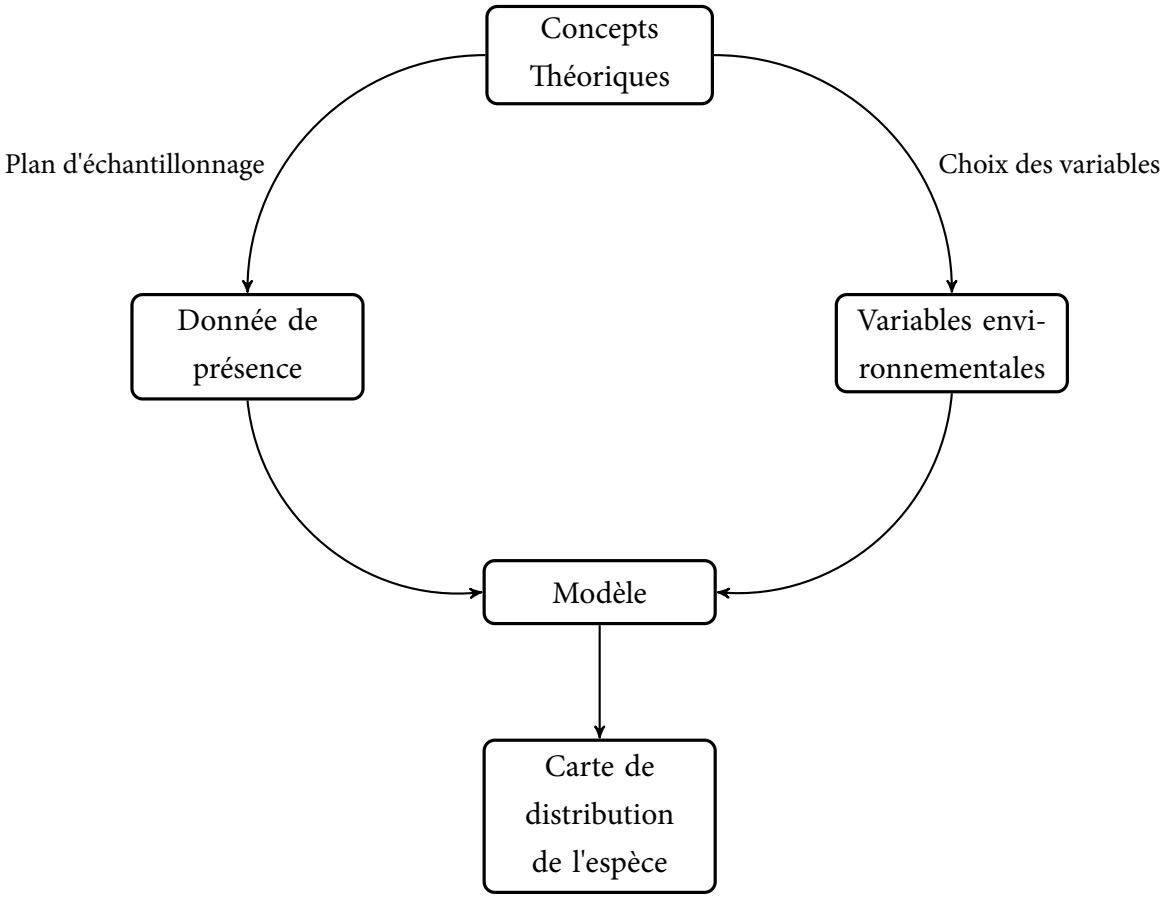
1. les conditions abiotiques comme l'altitude, le climat ou le couvert ;
2. les facteurs biotiques qui regroupent les différentes interactions que l'espèce entretient avec les autres espèces (i.e prédation, compétitions, etc.) ;

3. le capacité de dispersion spatiale de l'espèce, sa propension à pouvoir coloniser d'autres régions accessibles ;
4. la capacité d'adaptation de l'espèce, celle à résister au changement des milieux. Cependant, il s'agit d'une condition qui est souvent considérée comme négligeable.

Une question centrale est de savoir quel type de niche est modélisée dans les modèles de distributions d'espèces : est-ce la niche fondamentale ? la niche réalisée ? ou encore la probabilité d'utilisation de l'habitat ?
La réponse à cette question n'est pas tranchée et selon les auteurs et le type de modèle on distingue l'un ou l'autre de ces concepts. Pour Guisan et Thuiller (2005), les différents types de modèles, utilisés permettent de mettre en évidence la niche réalisée de l'espèce. La description de cette niche réalisée dans l'espace géographique représente la distribution potentielle de l'espèce ou encore l'habitat favorable. Cette distinction se révèle très importante lors de l'interprétation et l'intégration des résultats dans le cadre décisionnel.
Selon Hirzel et Le Lay (2008), la question ne se pose pas en pratique car il note que les interactions biotiques, parfois associées à une fragmentation de l'habitat ont souvent lieu à une résolution spatiale très fine, ce qui contraste avec les données souvent utilisées dans les études, dont l'échelle est souvent grossière (> 5km).

On peut aussi rajouter, que ces différents modèles ne peuvent cependant permettre une aide à la décision effective que si ces certaines conditions préalables sont respectées. Il s'agit principalement de conditions pratiques d'ordre écologique. Il faut que :

- Le système soit stable en termes de dynamique des populations et en flux migratoire ;
- l'habitat disponible (l'extent géographique) soit bien délimité ;
- la sélection d'habitat soit un choix, donc que l'accès de l'espèce à tous les habitats soit libre et égal ;
- les variables environnementales qui influencent la probabilité de sélection de l'habitat soient correctement identifiées ;



Graphique 1.3 : Principe des modèles de distribution d'espèce.

Pour la mise en place effective d'un modèle de distribution d'espèce (graphique 1.3), il faut :

- Établir un modèle conceptuel théorique sur les différents facteurs biotiques et abiotiques qui influencent la niche de l'espèce ;
- collecter des données d'occurrences en se basant sur des données de terrain (échantillonage), en utilisant des sources historiques ou par connaissance d'expert ;
- choisir les variables environnementales qui pourraient avoir une influence sur la distribution de l'espèce, en général il s'agit de carte digitale obtenue par télédétection ;
- choisir le ou les modèles appropriés selon les objectifs et l'expérience du modélisateur ;
- valider les modèles et interpréter les cartes finales.

1.3.6 Modèles sur la distribution des glossines

Les glossines font partie des arthropodes qui fascinent les chercheurs depuis plus d'un siècle. Les études sur leurs aires de répartition géographique à l'échelle du continent remontent aux travaux de W. H. Potts en 1954. Selon Rogers (2000), il existe deux approches pour modéliser la distributions des glossines : une approche prédictive (biologique) et

une approche descriptive (statistique). Les modèles dont se rapproche le plus le nôtre sont ceux développés dans les années 90, principalement par l'équipe Trypanosomaiasis And Land-use in Africa (TALA) du département de zoologie de l'université d'Oxford. Nous présentons ici quelques études qui ont essayé de prédire la distribution des glossines en utilisant des images de télédétection.

Rogers et Randolph (1991) montrent que la distribution et l'abondance des glossines sont étroitement liées aux conditions climatiques et que les images satellites peuvent alors être utilisées afin d'obtenir des indicateurs de ces paramètres climatiques. Ils trouvent une relation négative entre la mortalité des glossines et la température qu'ils approchent en utilisant des capteurs thermiques embarqués sur des satellites.

Rogers et Randolph (1993) utilisent des images du capteur Advanced Very High Resolution Radiometers (AVHRR) embarquées sur le satellite météorologique du National Oceanic and Atmospheric Administration (NOAA) et en particulier le NDVI (Normalised Difference Vegetation Indice, une mesure de l'activité photosynthétique de la végétation) pour prédire la distribution des glossines de l'espèce *morsitans* et *pallipides* au Kenya et en Tanzanie. Les données de terrain proviennent de sources historiques et ils arrivent à obtenir des pourcentages de bon classement autour de 80%.

Rogers *et al.* (1996) reprennent la même méthodologie que Rogers et Randolph (1993) et calculent la distribution de 8 espèces de glossines sur un transect allant de Bobo-Dioulasso (Burkina Faso) à Abidjan (Côte d'Ivoire). Ils utilisent en plus du NDVI, des indicateurs de température du sol et de pluviométrie tous dérivés d'images satellites. Cette analyse à été faite sur des pixels d'environ 28 km de côté, sur une surface d'environ 800 km. Ils utilisent une analyse discriminante et une régression logistique afin de produire des cartes de probabilité de présence. Ils remarquent par ailleurs dans cette zone de l'Afrique que les variables thermiques jouent un rôle plus important que celles liées à la végétation. En utilisant cette méthodologie, ils ont obtenu des cartes avec des précisions allant à jusqu'à 85% de bon classement sur des jeux de données de test. Cette méthodologie est à la base des cartes de risque du PAATEC, utilisées par la FAO.

Robinson *et al.* (1997a,b) utilisent une analyse discriminante, une classification basée sur le maximum de vraisemblance et une analyse en composantes principales afin d'établir la distribution de 4 espèces de glossines présente en Afrique Australe (ceinture commune de tsé-tsé). Les variables utilisées sont le NDVI, des mesures de température au sol et des données d'élévation et leurs analyses leurs permettent pour certaines espèces d'atteindre

des précisions de 92%.

Hendrickx *et al.* (1999a,b) utilisent des techniques de classification non supervisées et intègrent des images AVHRR dans leur méthodologie afin de réaliser des cartes du risque Trypanosomien. Dans leurs approches, ils utilisent aussi des variables liées aux facteurs biotiques comme les mouvements de troupeaux. Comme Rogers *et al.* (1996) ils trouvent une relation entre les densités de glossine et le NDVI. De plus ils mettent l'accent sur le rôle que peuvent jouer ces images satellites lors de campagne de collecte de données entomologiques.

Cecchi *et al.* (2008) réalisent dans une étude récente une cartographie de l'habitat des plusieurs espèces de glossine en Afrique. Leur approche se base sur une classification du type de sols qui utilisent le Land Cover Classification System (LCCS) à des résolutions spatiale différentes. Ils obtiennent des corrélations allant de 47% selon les classes de végétation du LCCS et les glossines du groupe *palpalis* à l'échelle continentale.

Chapitre 2

Présentation des données

Dans les modèles de distribution d'espèce, le choix des variables explicatives (données environnementales) et à expliquer (données de présence/absence) qui seront utilisées est une étape critique pour la réussite de l'analyse. Dans cette étude, on distingue deux types de données : les données de terrain et les données de télédétection. Ces données ont été collectées en se basant sur les connaissances théoriques dont on dispose sur *G. p. gambiensis*.

Les données de terrain ont été obtenues lors d'enquêtes entomologiques dans la zone d'étude. Les données de télédétection sont principalement constituées de séries temporelles d'images satellites qui permettent de suivre les différents phénomènes climatiques et environnementaux de la zone d'étude. Toutes ces données ont été retravaillées pour une utilisation plus effective dans les différents modèles qui seront présentés dans les chapitres qui vont suivre.

2.1 Télédétection et épidémiologie

La télédétection est la science et l'art d'obtenir des informations sur un objet, une aire ou un phénomène à travers l'analyse de données acquises par le biais d'un instrument qui n'est pas en contact direct avec l'appareil de mesure (Reddy, 2006).

L'épidémiologie est l'étude de la distribution des troubles de santé dans des populations humaines, animales ou végétales, de leurs facteurs de variation, et de leurs dynamiques spatiales et temporelles afin d'en déduire des mesures de surveillance et de contrôle. Elle s'intéresse, entre autres, aux relations entre des indices d'occurrences d'une maladie, d'une

infection ou d'un vecteur ou facteur de risque, et les caractéristiques de l'environnement.

Les images de télédétection, en particulier les images satellitales, permettent de capter plusieurs aspects du système biophysique de la surface de la terre. On dispose alors, par le biais d'image à haute résolution et de séries temporelles, d'informations sur des phénomènes environnementaux qui sont impossible à obtenir avec des mesures au sol. L'utilisation de données de télédétection en épidémiologie (Hay *et al.*, 1997, 2000) tient au lien étroit qui existe entre l'émergence de certaines maladies et les conditions climatiques. Ce lien est particulièrement fort pour les maladies transmises par des insectes vecteurs (moustiques, mouches, etc.) dont la distribution est liée à des paramètres tels que la température, l'humidité, qui déterminent leur type d'habitat (Tran, 2004).

Les progrès techniques des satellites, des capteurs et des traitements informatiques et l'importance de l'environnement dans l'émergence de nouvelles maladies ont entraîné le développement des applications de la géomatique à l'épidémiologie et l'importance de la télédétection dans la description, l'explication et la prédiction des maladies vectorielles est incontestable. Dans les études reliant la télédétection et la santé, de nombreux indices extraits d'images satellitales peuvent être utilisés pour caractériser l'environnement, l'objectif étant la recherche des facteurs de risque épidémiologique. Le lien entre le risque épidémiologique et la distribution des espèces vectrices de maladie est très fort. En effet, l'occurence du vecteur est l'un des indicateurs de risque les plus important en épidémiologie (Tran *et al.*, 2005). Nous présentons dans cette section les données dérivées des images qui seront utilisées pour la caractérisation de la distribution de *G. p. gambiensis* dans la zone géographique des Niayes. L'utilisation potentielle qui peut être faite de ces images dépend en grande partie de certaines propriétés inhérente à ce type de donnée comme leurs *résolutions spatiales, spectrales et temporelles*.

2.1.1 Résolutions spatiale, spectrale et temporelle

Résolution spatiale

La résolution spatiale d'une image est liée à la taille des pixels constituant l'image. Cette résolution dépend du type d'appareil utilisé. La taille varie de l'ordre de quelques centimètres pour des satellites civils (61cm pour des images QuickBird) à plusieurs kilomètres pour les satellites météorologiques. Cependant, la zone couverte (fauchée) est plus importante lorsque la résolution est moins fine. La fauchée d'une image QuickBird par exemple est de 16,5 km contre 3000 km pour une image provenant du satellite météorologique NOAA.

Résolution temporelle

La résolution temporelle ou cycle orbital correspond à la fréquence avec laquelle un satellite va pouvoir acquerir la même scène. Les résolutions spatiales et temporelles sont inversement liées : une résolution temporelle fine (répétitivité importante) ne pourra être obtenue que pour des images à basse résolution spatiale, et inversement.

Résolution spectrale

Les ondes du spectre électromagnétique se différencient selon leur longueur. En télédetection, le rayonnement éléctromagnétique est le plus souvent dans le visible, le proche, le moyen infrarouge, l'infrarouge thermique et le radar (domaine des hyperfréquences, qui traversent l'atmosphère). Les images peuvent être acquises avec des capteurs optiques. L'objet étudié détermine l'information spectrale à mesurer. A titre d'exemple, l'occupation du sol peut être mesurée par des capteurs optiques, la température de la surface de la terre par des capteurs thermiques et la pente ou l'humidité par des capteurs radars. Il est parfois nécessaire de mesurer le rayonnement dans plusieurs longueurs d'onde pour identifier un objet ou un paysage (Guis, 2007). On obtient ainsi un profil appelé *signature spectrale* propre à l'objet.

Obtenir des images de bonne résolution temporelle et spatiale semble être une tâche difficile mais nécessaire à l'étude de phénomènes complexes. De plus, une grande résolution spectrale permettrait d'avoir une large gamme de mesure des phénomènes climatiques et environnementaux. Les récents progrès faits dans le domaine de la télédetection permettent aujourd'hui de disposer de tels produits : les images du spectroradiomètre imageur à résolution moyenne (MODIS).

2.2 Transformation de Fourier des données MODIS

2.2.1 Moderate Resolution Imaging Spectroradiometer : MODIS

Le spectroradiomètre imageur à résolution moyenne (MODIS) est embarqué sur les satellite Terra (lancé en 1999) et Aqua (lancé en 2002) de la NASA. Ces deux satellites sont complémentaires : le premier, Terra, fait son orbite autour de la terre deux fois par jour, il passe à l'équateur à 10h30 et 22h30 (heure solaire locale) alors que le satellite Aqua passe en début d'après midi à 13h30 et à 01 :30 (Wan, 2006). Nous avons alors la possibilité d'avoir quatre images par jour en utilisant les images MODIS fournies par ces satellites.

Parmi les différents capteurs et types d'images satellites utilisées, les images MODIS sont considérées comme ayant un bon compromis entre résolution spatiale et temporelle et c'est une des raisons pour lesquelles elles sont grandement utilisées dans les études épidémiologiques (Hay *et al.*, 1997, 2000).

Nous disposons d'environ 11 ans d'images provenant du satellite Terra et 8 ans du satellite Aqua. Les capteurs de ces satellites fournissent gratuitement et librement des images MODIS à haute résolution temporelle et moyenne résolution spatiale (250m à 1km). Toutes ces séries temporelles d'images satellites permettent de calculer plusieurs indices. Les indices calculés permettent de quantifier la dynamique des éléments de l'environnement et du climat.

Indices de température

La température est un paramètre qui joue un rôle important dans le cycle de vie de glossines et parmi les indicateurs de température, le Land Surface Temperature est l'un des plus utilisé. Le Land Surface Temperature (LST) est calculée à partir de la mesure des radiations émises par la surface de la terre. L'algorithme du LST est basé sur la loi de Planck qui permet de relier la radiance émise par un corps noir [1] à travers des longueurs d'onde précises à une certaine température (Wan, 1999). Il s'agit d'un indicateur très corrélé à la température de l'air (Vancutsem *et al.*, 2010), et il est utilisé dans beaucoup d'études de distribution d'espèce et d'épidémiologie spatiale (Neteler, 2010).
Dans ce document, nous utiliserons deux indicateurs pour avoir une mesure de la température de l'air : le *LST day* et le *LST night*. Le premier est une approximation de la température moyenne en journée et le second mesure l'activité thermique moyenne pendant la nuit.

Indices de végétation

Parmi les indices couramment utilisés dans les études épidemiologiques, le NDVI (Normalized Difference Vegetation Index) est une mesure de l'activité chlorophyllienne. Cet indice permet de distinguer les sols nus de la végétation, il est étroitement corrélé avec le pourcentage de recouvrement d'occupation du sol par un couvert végétal. En plus du NDVI, il existe d'autres indices de végétation comme le EVI (Enhanced Vegetation Index) mais selon Hay *et al.* (1996), le NDVI est particulièrement utile dans des zones où la végétation est peu dense alors que dans les zones forestières on favorisera plutôt l'EVI. Vu le type de végétation présente dans la zone des Niayes, le NDVI est le plus approprié et sera donc utilisé pour capter la vigueur de la végétation arborée dans notre zone d'étude. On note aussi qu'il s'agit d'un indicateur qui a déja été utilisé à plusieurs reprises pour prédire la densité de glossines (Rogers *et al.*, 1996 ; Guerrini, 2009) en Afrique de l'Ouest.

1. un corps qui absorbe toutes les radiations à n'importe qu'elle longueur d'onde et qui n'en reflète aucune

Indices de réflectance

La réflectance dans le moyen infrarouge (MIR) permet de mesurer le rayonnement des sols nus. Cet indice est corrélé à la température de la surface de la terre et de la structure des canopées [2] (Boyd et Curran, 1998). Cependant, par rapport aux autres indicateurs, on note qu'il est peu sensible aux interférences atmosphériques. Une végétation luxuriante est caractérisée par un faible MIR. Cet indice permet en plus du NDVI de caractériser la végétation, mais aussi la température du sol.

En plus de ces indicateurs dynamiques, nous utiliserons un modèle numérique de terrain qui permet d'avoir des données d'élévation sur notre zone d'étude.

Pour cette étude, nous disposons d'un indicateur statique (l'élevation) et de quatre séries temporelles d'indices liés au climat et l'environnement : le LST day, LST night, le NDVI et MIR. Pour ces indicateurs dynamiques, on dispose de données qui vont du 1^{er} janvier 2001 au 31 janvier 2008, et pour chaque année, on a 46 images (une observation tous les 8 jours) et toutes les images (y compris l'élévation) sont à une résolution spatiale de 1 km. Deux types de produits MODIS ont été utilisés MOD11A2 et MCD43B4 pour calculer ces différents indices avec une résolution spatiale d'environ 1 km (voir tableau 2.1). Nous disposons alors de plusieurs séries temporelles qui nous permettent finalement d'avoir 8 ans de données sur le climat et la végétation dans la zone des Niayes.

Tableau 2.1 : Liste des différents produits MODIS utilisés.

produits	résolution spatiale	résolution temporelle	indicateurs
MOD11A2	1000 m × 1000 m	8 jours	LST day (LSTD) et night (LSTN)
MCD43B4	1000 m × 1000 m	8 jours	EVI, NDVI, MIR

2.2.2 Analyse de Fourier temporelle des images MODIS

Les différents indicateurs présentés ci-dessus permettent de capter la dynamique des changements des caractères abiotiques dans notre zone d'étude. Ils sont donc très importants pour la modélisation de la distribution de *G. p. gambiensis*. La haute résolution temporelle des images MODIS est une mine importante et utile d'informations. Cependant, en tant que série temporelle, elle présente l'inconvénient d'une auto-corrélation élevée entre observations. De plus, même avec des images composites par 8 jours, le nombre d'observation est très grand sur une longue période d'étude : 365 données pour la période 2001-2008.

2. étage supérieur de la forêt, qui dépend des rayonnements solaires

Afin de contourner ces différents problèmes, plusieurs méthodes ont été proposées, dont la plus populaire est l'analyse en composantes principales (ACP). Mais cette technique, si elle permet une compression de l'information efficace, souffre d'un problème majeur : *on perd toute trace de la dynamique temporelle et des changements saisonniers*. Dans leur modèle de distribution des glossines en Afrique de l'Ouest, Rogers *et al.* (1996) proposent une méthodologie originale qui permet une compression de l'information aussi efficace que celle réalisée par une ACP classique, tout en gardant la dynamique temporelle : leur solution est d'utiliser une transformation de Fourier (Bloomfield, 2004) des séries temporelles d'images.

La transformation de Fourier est un cas particulier de l'analyse harmonique [3], elle permet donc de passer du paradigme *temporel* à celui des *fréquences*. Ainsi, les séries temporelles originales sont transformées en somme d'harmoniques (sinusoïdes) de fréquences, phases et amplitudes différentes. En plus de retranscrire par le biais d'une phase et d'une amplitude chaque signature temporelle (attachée à une fréquence donnée), les différentes harmoniques ont la propriété d'être orthogonales (non corrélées entre elles). De plus, la somme de toutes les harmoniques permet de retrouver toutes les variations présente dans la série originale [4]. La méthodologie originale de Rogers *et al.* (1996) à été largement amélioré par Scharlemann *et al.* (2008), en utilisant un algorithme itératif basé sur une interpolation des données par splines. Une des propriétés des séries temporelles dans l'espace des fréquences, est qu'un signal de fréquence élevée correspond à une faible fluctuation de la série originale. Cette fréquence est dès lors associée à un « bruit » ou une perturbation . Donc en plus de pouvoir capter la saisonnalité présente dans la série de base, la Temporal Fourier Analysis (TFA) permet en outre de lisser les séries temporelles d'images. Nous donnons, les bases du calculs de Fourier qui ont été nécessaires pour transformer les différentes séries de données. Une explication détaillée de l'algorithme se trouve dans Scharlemann *et al.* (2008).

Soit x_t une série temporelle de N observation qui représente par exemple l'évolution d'un indicateur (e.g LST day) climatiques sur un pixel d'image MODIS (pixel de 1 km). La TFA permet de réaliser une décomposition de x_t :

$$x_t = a_0 + \sum_{p=1}^{\frac{N}{2}-1} \left[a_p \cos\left(\frac{2\pi t}{N}\right) + b_p \sin\left(\frac{2\pi t}{N}\right) \right] + a_{N/2} \cos(\pi t), \left(t = 1, 2, \cdots, N\right)$$

où les coefficients (a_0, b_0) sont définis par :

3. projection dans une base sinusoïdale
4. à travers une transformation de Fourier inverse

$$a_0 = \bar{x}$$
$$a_{\frac{N}{2}} = \frac{\sum(-1)^t x_t}{N}$$
$$a_p = \frac{2(\sum x_t \cos(\frac{2\pi pt}{N})}{N}$$
$$b_p = \frac{2(\sum x_t \sin(\frac{2\pi pt}{N})}{N}$$

2.2.3 La notion d'amplitude et de phase

Considérons la p^{ieme} harmonique de fréquence $\omega p = \frac{2\pi p}{N}$. Il est alors possible d'obtenir une décomposition de la série temporelle d'origine en plusieurs harmoniques de fréquences différentes. Chaque harmonique peut être écrite comme :

$$a_p \cos \omega_p t + b_p \sin \omega_p t$$

En utilisant les identités trigonométriques classiques, on peut réecrire cette harmonique :

$$R_p cos(\omega_p t + \phi_p) = a_p \cos \omega_p t + b_p \sin \omega_p t$$

On remarque alors que pour chaque fréquence ω_p, les harmoniques sont définies par deux valeurs : R_p et ϕ_p. La première valeur, R_p, est l'amplitude de la sinusoïde et on a :

$$R_p = \sqrt{(a_p^2 + b_p^2)}$$

Quant à la seconde valeur, ϕ_p, elle s'appelle la phase et elle représente la position de l'harmonique lorsque cette dernière atteint son maximum. On a par identification :

$$\phi_p = tan^{-1}(\frac{-b_p}{a_p})$$

Ces deux mesures complémentaires permettent alors pour chaque fréquence de décrire les mouvements saisonniers de la série originale. Mais elles dépendent entièrement de la fréquence. Comment choisir alors les fréquences qui permettent décrire fidèlement la série originale ?

2.2.4 Choix des fréquences

LA TFA permet de décomposer la série temporelle en somme d'harmoniques décorrélées de fréquences différentes. Toutes les fréquences ne décrivent pas les mêmes fluctuations de la série originale. Il est alors nécessaire de ne garder que les harmoniques associées

aux fréquences qui explique la plus grande part de la variance de la série originale. Le résultat ci-dessous permet d'avoir la contribution de chacune des harmoniques à la variance totale : c'est une implication du théorème de Parseval (Rudin, 1987). On a pour les $\frac{N}{2}$ harmoniques :

$$\frac{1}{N}\sum(x_t - \bar{x})^2 = \frac{1}{2}\sum_{p=1}^{\frac{N}{2}-1} R_p^2 + a_{\frac{N}{2}}^2$$

On peut remarquer que $R_p^2/2$ est la contribution de la de la p^{ieme} harmonique à la variance totale de la série temporelle d'origine.
Pour les transformées de Fourier utilisées avec les images MODIS, les trois premières harmoniques correspondent aux cycles annuel, semestriel et quadrimestriel. En général, ce sont les seuls fréquences extraites (Scharlemann *et al.*, 2008).
Le cycle qui explique la plus grande part de la variance pour chacun de nos indicateurs dynamiques (LST, NDVI, MIR) est le cycle annuel. Donc pour le reste de l'analyse la seule harmonique utilisée, sera celle correspondant au cycle annuel. Cette harmonique sera alors décrite par son amplitude et sa phase.

Variables environnementales et climatiques utilisées

Nos indicateurs dynamiques ont été transformées à l'aide d'une TFA. En plus de la phase et de l'amplitude du cycle annuel, nous utiliserons aussi la moyenne, le minimum et le maximum de chaque variable sur 8 ans. De plus, nous utiliserons aussi des données d'élévation sur la zone d'étude.

Tableau 2.2 : Variables environnementales et climatiques disponibles

Nom	Description	Unité
LSTDa0	LST day moyen entre 2001 et 2008	°C
LSTDa1	LST day amplitude du cycle annuel	°C
LSTDp1	LST day phase du cycle annuel	*mois de l'année*
LSTDmn	LST day minimum entre 2001 et 2008	°C
LSTDmx	LST day maximum entre 2001 et 2008	°C
LSTNa0	LST night moyen entre 2001 et 2008	°C
LSTNa1	LST night amplitude du cycle annuel	°C
LSTNp1	LST night phase du cycle annuel	*mois de l'année*
LSTNmn	LST night minimum entre 2001 et 2008	°C
LSTNmx	LST night maximum entre 2001 et 2008	°C
NDVIa0	NDVI moyen entre 2001 et 2008	sans unité
NDVIa1	NDVI amplitude du cycle annuel	sans unité
NDVIp1	NDVI phase du cycle annuel	*mois de l'année*
NDVImn	NDVI minimum entre 2001 et 2008	sans unité
NDVImx	NDVI maximum entre 2001 et 2008	sans unité
MIRa0	MIR moyen entre 2001 et 2008	sans unité
MIRa1	MIR amplitude liés la fréquence annuelle	sans unité
MIRp1	MIR phase du cycle annuel	*mois de l'année*
MIRmn	MIR minimum entre 2001 et 2008	sans unité
MIRmx	MIR maximum entre 2001 et 2008	sans unité
elev	modèle numérique de terrain, élévation	*m*

source : TALA, Oxford

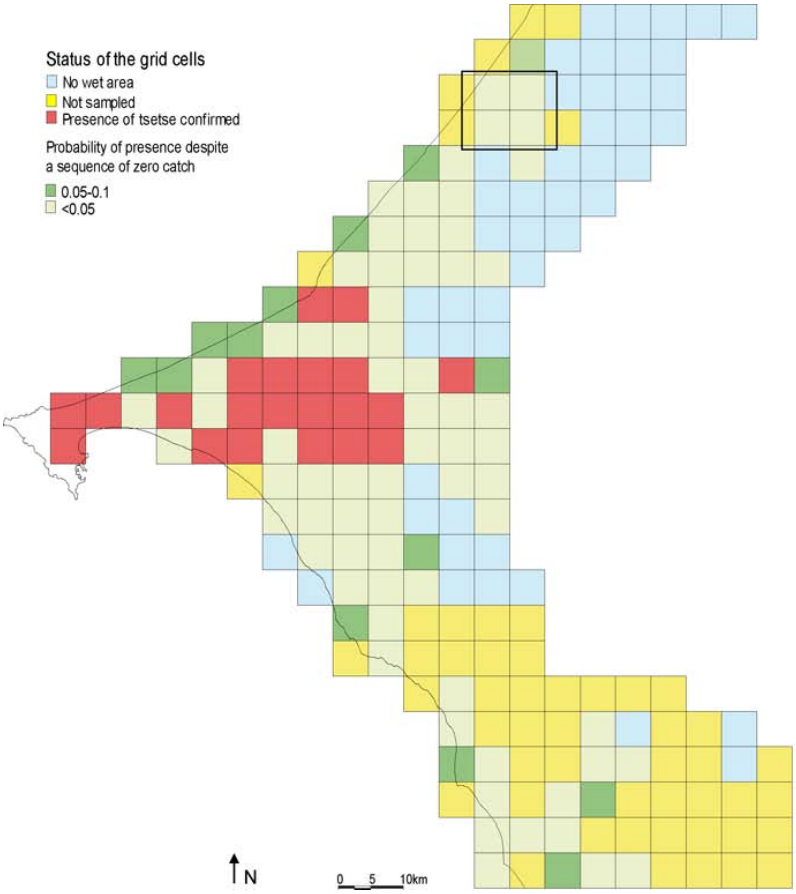
Le tableau 2.2 donne l'ensemble des variables qui sera utilisées dans les différents modèles de niche écologique et de prédiction de l'habitat favorable pour *G. p. gambiensis* dans les Niayes. On remarque que les phases ont été converties en mois pour en faciliter l'interprétation biologique, pour les variables de température elle sont en degré celcius, la réflectance dans le moyen infrarouge et le NDVI sont sans unité. Ces deux derniers indicateurs (MIR, NDVI) sont compris entre -1 et 1.

2.3 Données de terrain et délimitation de la zone d'étude

Les données de télédetection décrites plus haut seront mis en relation avec des données de terrain par le biais de modèles spécifiques. Ces données de terrain, sont de sources diverses mais les principales données sont celles de présence et d'absence.

2.3.1 Données de présence et d'absence

Les données de présence et d'absence qui seront utilisées dans ce document, sont issues d'une campagne d'échantillonnage stratifié à travers la zone des Niayes (Bouyer *et al.*, 2010b). Bouyer *et al.* (2010b) ont réalisé cette enquête entomologique en utilisant des images LandSat (Land Satellite). Ces images ont permis de réaliser une classification du couvert végétal de la zone des Niayes. Cette approche basée sur une enquête phytosociologique préalable, a permis d'obtenir les principales *zones qui restent humides durant la saison sèche* qui représentent les seuls endroits où l'on peut trouver l'espèce à cette période de l'année. Cette méthode a permis de restreindre la zone d'échantillonnage à 4% de la surface totale. Afin de vérifier la qualité des données de présence et d'absence sur les différents pixels, un modèle probabiliste basé sur l'efficacité des pièges à tsé-tsé à été établi (Barclay et Hargrove, 2005) sur des pixels de 5 km de largeur. Il permet de donner la probabilité qu'une mouche soit présente sur le pixel étudié malgré le fait qu'on ait une série de données d'absence sur ce pixel.



Graphique 2.1 : Plan d'échantillonnage sur la zone d'étude : les pixels rouges représente les zones où la présence de l'espèce a été confirmée, en jaune les zones non échantillonnées, en bleu il s'agit de zones non échantillonnées et défavorables aux glossines en saison sèche. Reproduit avec la permission de Bouyer *et al.* (2010b)

Toutes les zones n'ont pas été échantillonnées (graphique 2.1), et le choix de la saison sèche pour l'enquête facilite les opérations car l'aire de répartition des glossines riveraines est restreinte aux zones de végétation humide. Nous avons constitué un jeu de données sélectionné de façon raisonnée, en nous appuyant sur les connaissances disponibles sur l'écologie de *G. p. gambiensis*. Nous avons ainsi écarté la couche d'information décrivant les zones humides de saison sèche, car elle avait été obtenue en 2001, bien avant l'enquête entomologique : les zones avaient pu changer au gré des aménagements agricoles et des changements de l'occupation des sols.

2.3.2 Données d'absences

Dans les différents modèles de distribution d'espèces décrits au chapitre 1 une grande partie utilise pour la modélisation à la fois les absences et les présences (e.g régression logistique, forêt aléatoire, etc.) et d'autres modèles utilisent principalement les données de

présence. Les données d'absences sont par nature différentes des données de présence, car de multiples raisons peuvent conduire à l'absence observée après une ou plusieurs séances de capture :

1. l'espèce n'a pas été détectée par le piège alors que l'espèce est présente dans la zone de piégeage (efficacité du piège, contrainte spatio-temporelle, etc.) ;
2. l'espèce n'est pas détectée dans une zone favorable pour des raisons historiques comme le cas de campagnes d'éradication ;
3. l'espèce n'a pas été détectée car la zone est vraiment défavorable à l'espèce.

Le but de cette étude est d'obtenir une estimation de la niche potentielle *G. p. gambiensis* donc seul la troisième raison pour laquelle on a une absence nous est utile. Dans le but de corriger ces problèmes liés à la nature des données d'absences nous allons utiliser un modèle probabiliste développé par Barclay et Hargrove (2005). Ce modèle se base sur l'efficaté des pièges utilisés lors de la collecte des données. L'utilisation de ce modèle à une échelle plus fine que celle utilisée pendant l'enquête entomologique de Bouyer *et al.* (2010b) et le choix d'un seuil beaucoup plus sévère, nous a permis de réduire de 38% le nombre de d'absence sur la zone d'étude. Dans le reste de l'analyse la probabilité qu'un pixel déclaré comme non infesté ne le soit pas est inférieur à 10^{-6}. Ces absences permettent de résoudre la première raison pour lesquelles les absences sont entachées d'incertitudes.

2.3.3 Données auxiliaires

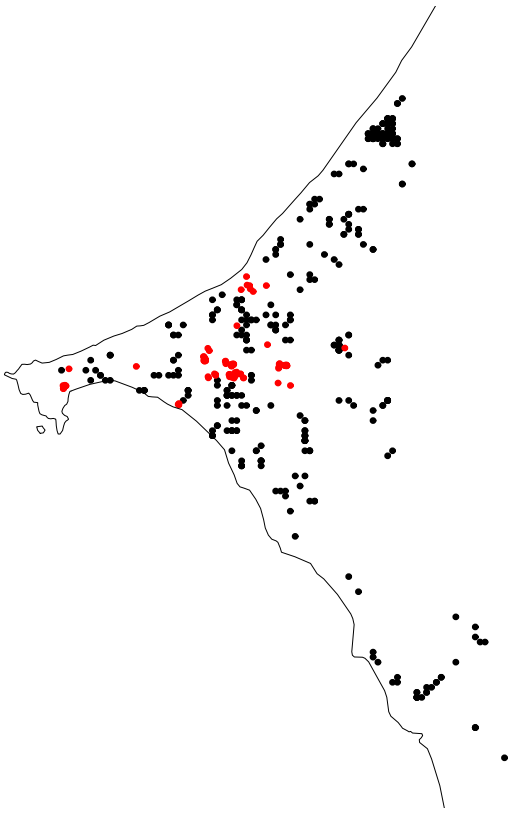
En plus de l'enquête entomologique, nous disposons aussi des données issues de la campagne phytosociologique menée préalable a la classification de la zone d'étude. Ces données permettent de classer certaines zones comme des gîtes potentiels pour *G. p. gambiensis*. Il s'agit d'une classification à dire d'expert sur des sites choisis minutieusement. Les sites ont été séparés en trois classes : les sites favorables, les sites défavorables et les sites favorables mais dégradé. Ces données seront utilisées dans la validation qualitative des différents modèles de prédictions.

2.3.4 Choix de la délimitation de l'espace géographique

Dans l'étude des modèles de distribution d'espèces, certains choix importants doivent être pris par le modélisateur et le biologiste, il s'agit de paramètres importants qui ont une grande influence sur les résultats. Parmi ces choix, deux sont critiques et d'une grande importance : *la délimitation de la zone d'étude* et *l'échelle ou la résolution spatiale* .

Choix de la délimitation de l'espace géographique

Dans les modèles de distributions d'espèce, à part les modèles de présence-seule, tous les autres types de modèles sont extrêmement sensibles aux choix de la zone d'étude. Cette sensibilité est particulièrement vraie pour les modèles de présence-background. Au Sénégal, la zone du Siné-Saloum au sud de la zone d'étude est connue pour abriter une grande population de *G. p. gambiensis* mais Bouyer *et al.* (2010b) ont montré à travers une étude de génétique des populations que la population de glossines du sud et celle des Niayes sont génétiquement différents. En conséquence, nous avons considéré que la population de *G. p. gambiensis* présente dans les Niayes était isolée des autres population de la même espèce. Notre analyse a porté sur cette population particulière qui a réussi à s'adapter à des conditions environnementales éloignées de celles où elle est habituellement rencontrée.



Graphique 2.2 : Données de présence (rouge) et d'absence (**noir**) sur la zone des Niayes.

Choix de la résolution spatiale

Le choix de la résolution spatiale est généralement déterminé par la résolution spatiale des données raster (image) dont on dispose. Plus la résolution est fine, et plus la modélisation sera précise. Notons que dans le cadre de cette étude, la résolution spatiale sera de 1 km, ce qui représente la résolution de nos produits MODIS [5].

5. il existe cependant des produits MODIS de résolution plus fine allant jusqu'à 250 m

Chapitre 3

Analyse exploratoire de la niche

Parmi les différents types de modèles de distribution, deux types d'études peuvent être entreprises afin de comprendre la relation entre l'espèce et son environnement. D'un côté, nous avons les analyses exploratoires dont le but est de trouver dans un grand ensemble de variables environnementales, celles qui jouent un rôle important dans la distribution de l'espèce. Et d'un autre côté, les modèles prédictifs dont le but est de prédire la distribution potentielle des espèces. Mais les différents modèles prédictifs sont souvent basés sur des modèles de régression qui sont en général sensibles au ratio nombre de variables sur nombre d'observations. Ces modèles ne sont donc efficients en général qu'avec un nombre limité de variables. Des méthodes purement quantitatives sont souvent utilisées dans le but de réduire le nombre de prédicteurs, et ces méthodes sont basées sur des critères comme les critères d'informations classiques (Akaike Information Criterion, Bayesian Information Criterion, etc.), ou encore des modèles de régression pas à pas (stepwise regression). Ces critères ne se concentrent principalement que sur deux points : le *pouvoir prédictif* et la *parcimonie*. Donc dans ces différentes procédures, aucun critère ne nous assure de garder les variables qui jouent un rôle important dans l'écologie de l'espèce étudiée Meynard et Quinn (2007). Nous favorisons donc une nouvelle approche basée sur *l'interprétabilité scientifique* des différents modèles (Cressie et Wikle, 2011). Il devient alors nécessaire pour le modélisateur d'avoir des outils lui permettant de réaliser une exploration d'un grand nombre de variables afin de connaître et de comprendre les facteurs qui peuvent influencer la distribution de l'espèce étudiée. Ces outils, dans le cadre des concepts mis en place, permettront de construire alors un cadre d'analyse cohérent pour l'élaboration de modèles prédictifs.

Une implication directe de la théorie de la niche écologique (Hutchinson, 1957) est la

conceptualisation du système écologique de l'espèce sous une forme analytique et géométrique. En effet, chaque variable environnementale définit une dimension dans un espace (multidimensionnel) écologique. Cette formalisation de l'espace écologique permet de réaliser des analyses quantitatives et graphiques et une exploration de cet ensemble dans le but d'identifier les variables qui jouent un rôle important dans le choix de l'habitat et la délimitation de la niche de l'espèce. Cependant, par essence, il s'agit d'un phénomène multidimensionnel dont l'exploration est difficile dès que nous dépassons le cadre d'un espace à trois dimensions. Alors comment résumer l'information que recèle la représentation multidimensionnelle de la niche de l'espèce étudiée ?
Les méthodes exploratoires multidimensionnelles classiques (Lebart *et al.*, 2008) sont des candidats naturels pour l'exploration de tels ensembles de données, en particulier les méthodes factorielles purement descriptives. Les analyses factorielles nous donnent la possibilité d'analyser de grands espaces multidimensionnels et d'en tirer le maximum d'information. Une adaptation de ces méthodes à l'espace écologique semble alors être une solution à l'exploration effective la niche de l'espèce.
Dans ce chapitre, nous présenterons de telles méthodes et finalement nous les appliquerons à la niche de *G. p. gambiensis* afin d'obtenir les variables les plus pertinentes dans la description de l'habitat de cette espèce. Le but final de ces analyses exploratoires est de permettre de construire un modèle conceptuels du système biologique étudié.

3.1 Méthodologie : introduction des méthodes factorielles

Les méthodes factorielles sont très anciennes et sont des outils de choix pour résumer l'information en présence de grand espace multidimensionnel. La méthode factorielle la plus ancienne et la plus utilisée est l'Analyse en Composante Principale (ACP) (Lebart *et al.*, 2008). Il s'agit d'une analyse qui permet une réduction effective des données contenues dans des grands tableaux, en utilisant une approche géométrique basée sur la notion d'inertie.

L'utilisation de ces méthodes dans le cadre de la distribution des espèces est assez récente (Hirzel *et al.*, 2002 ; Basille *et al.*, 2008 ; Calenge *et al.*, 2008) et repose principalement sur deux méthodes complémentaires : L'Analyse Factorielle de la niche Ecologique (ENFA) et la décomposition factorielle de la distance de Mahalanobis (MADIFA).

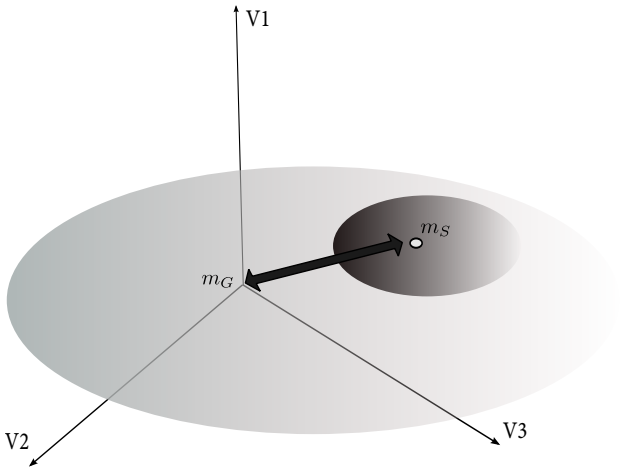
3.1.1 Ecological Niche Factor Analysis : ENFA

L'ENFA (Hirzel *et al.*, 2002) est une méthode d'analyse factorielle très proche de l'analyse en composante principale (ACP). L'originalité de cette méthode par rapport à une ACP

classique vient du fait que les axes sont construits afin d'avoir un sens écologique clair. Et la construction de ces axes est basée sur deux concepts importants : la *marginalité* et la *spécialisation*.
Comme en ACP, le but est de rechercher des directions (des axes, plans) qui permettront de réaliser une photographie [1] optimale de l'espace écologique dans un plan typiquement formé d'un axe de marginalité et d'un axe de spécialisation. Dans la description de cette méthode, nous utiliserons l'approche de Basille *et al.* (2008).

Le concept de marginalité

La marginalité peut être considérée comme une mesure de position au même titre qu'une moyenne, une médiane ou un mode en statistiques descriptives classiques. Elle mesure le carré de la distance entre l'espace disponible moyen, et l'espace moyen utilisé par l'espèce. D'un point de vue géométrique il s'agit de la norme entre l'origine de l'espace écologique et le centre de gravité de l'espace utilisé par l'espèce.



Graphique 3.1 : Marginalité définie comme la norme du vecteur reliant m_g à m_s, l'espace disponible en gris clair et l'espace utilisé en gris foncé.

L'habitat disponible (zone gris clair sur le graphique 4.7) est décrit par un ensemble de p variables environnementales. Chaque variable est associée à autant de cartes de type raster de la zone entière, elles sont constituées chacune de n pixels.

Supposons que X soit une matrice de $M_{n,p}$ contenant les valeurs de p prédicteurs (cartes) sur les n pixels disponibles. Dans l'espace écologique, X définit alors un nuage de points.
Afin de faciliter les démonstrations qui vont suivre, on suppose en plus que les différentes colonnes de notre matrice sont standardisées (de moyenne nulle et variance unité) afin

1. projection

que le centroïde soit situé à l'origine du repère de l'hypervolume définit par les différentes variables. Soit m_g le barycentre de ce nuage de points.
Comme tous les pixels ne sont pas utilisés par l'espèce, on peut définir un vecteur de taille n qui informe sur l'utilisation de chaque pixel par l'espèce. Il s'agit en pratique de compter le nombre de pièges par pixel où l'on a enregistré des captures de glossine. On peut aussi transformer ces données de comptage en *poids* [2]. Ce vecteur qu'on note ω est une mesure de l'utilisation de l'espace par l'espèce. Quand les éléments de ω sont supérieurs à zéro, elles correspondent à la zone gris foncée sur le graphique 4.7, on note D_p la matrice contenant ces poids en diagonale et m_s le centroïde du nuage de points. Ce centre de gravité, représente une utilisation moyenne de l'habitat. On note 1_n le vecteur colonne unité (des 1 partout), on note D la matrice diagonale contenant les poids pour les pixels disponibles, par défaut on a :

$$D = Diag\left(\frac{1}{n}\right)$$

.

$$m = X^\top D_p 1_n$$

La marginalité globale M est donc la norme du vecteur m.

$$M = m^\top m$$

Dans la suite de l'analyse, le vecteur de marginalité est normé :

$$q = \frac{m}{\sqrt{m^\top m}}$$

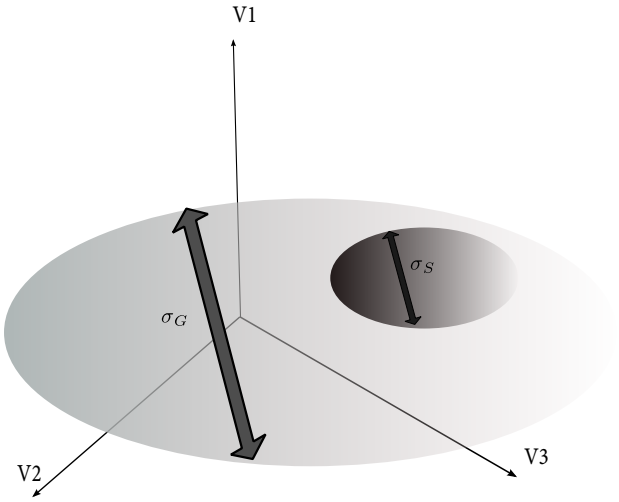
La marginalité est donc un indicateur de position qui mesure la déviation de la niche de l'espèce par rapport à l'espace disponible. Donc plus elle est grande et plus l'espèce préfère des conditions qui s'écartent fortement des conditions moyennes disponibles dans la zone d'étude.

Le concept de spécialisation

La spécialisation est une mesure complémentaire à celle de marginalité, de la même façon qu'une mesure de dispersion l'est par rapport à une mesure de position. Elle donne la forme de la niche en mesurant son degré d'étroitesse. Il s'agit du rapport entre l'inertie du nuage de point disponible sur l'inertie du nuage de point utilisé. D'un point de vue formel, il s'agit de chercher dans l'espace orthogonal à celui engendré par le vecteur de marginalité $p - 1$ facteurs qui maximisent le ratio des inerties entre les nuages : celui disponible et

2. soit ω_i ce poids alors $\sum_i^n \omega_i = 1$

celui utilisé (graphique 3.2). Une spécialisation forte dans une direction (i.e une variable) de l'espace écologique implique que la variance de nuage de l'espace disponible est grand par rapport à celle de l'espace utilisé, donc la niche est *étroite* par rapport à ce qui est disponible à l'espèce pour cette variable. Il s'agit alors d'une mesure unidimensionnelle (sur une direction précise).



Graphique 3.2 : Spécialisation, définie comme le rapport des variances entre les deux nuages : celui disponible et celui utilisé.

De manière analytique il s'agit de trouver un vecteur u tel que :

$$Max \frac{(Xu)^\top D(Xu)}{(Xu)^\top D_p(Xu)}$$
$$u^\top u = 1$$
$$u^\top m = 0$$

On peut montrer (Hirzel *et al.*, 2002) que ce programme de maximisation permet d'obtenir des vecteur u sur lequel la spécialisation est maximale. En résumé, il s'agit donc de maximiser le rapport entre les inerties des deux nuages dans un espace orthogonale à celui engendré par le vecteur de marginalité. On note que par construction les axes de spécialisations ne sont pas orthogonaux entre eux (démonstration en annexe A), mais par contre ils le sont par rapport à l'axe de marginalité. Donc pour pouvoir représenter les variables ou les individus (pixels) sur un plan factoriel issu de cette analyse, on utilise l'un des $P - 1$ plans constitués par l'axe de marginalité et un des axes de spécialisation.

Étapes de l'algorithme

En résumé, le principe de l'ENFA est le suivant :

– On calcule d'abord le vecteur de marginalité ;

– On projette le nuage de l'espace écologique sur le plan orthogonal au vecteur de marginalité ;
– On cherche les directions dans ce sous-espace (de dimension $p-1$) où la spécialisation est la plus grande possible, afin d'obtenir les différents axes de spécialisations ;

Interprétation des axes et liens avec la niche écologique

L'ENFA permet de réaliser une projection optimale de la niche dans un plan. Donc le premier plan représenté par l'axe de marginalité et le premier axe de spécialisation est la meilleure approximation possible de la niche réalisée sur un plan. L'interprétation se fait graphiquement en utilisant le biplot [3] (Gabriel, 1971), ce qui permet de donner directement l'influence des variables sur le choix de l'habitat par l'espèce. Les variables représentées par des vecteurs de normes élevées sont celles qui jouent un rôle critique dans le choix de l'habitat. L'angle réalisé par ce même vecteur dans le plan est une mesure du degré de marginalité ou/et de spécialisation, plus cette angle est faible ou plat (proche de l'axe des abscisses) et plus cette variable est caractéristique de la marginalité, d' autre part plus l'angle est droit (proche de l'axe des ordonnées) et plus la variable représente un facteur de spécialisation chez l'espèce.
On remarque alors que les coordonnées de chaque variable m_i sur le premier axe représentent le degré de marginalité qu'engendre cette variable. Plus cette valeur est élevée en valeur absolue, plus l'espèce se démarque des conditions moyennes telles que définies par l'environnement disponible. De plus un coefficient positif implique une préférence pour des valeurs de cette variable supérieure à la moyenne disponible pour la même variable. De même un coefficient négatif entraîne une préférence pour des valeurs plus faibles que la moyenne.
Quant à l'axe de spécialisation, l'interprétation est ici quelque peu différente de celle de l'axe de marginalité. La première différence est que le signe ne joue aucun rôle dans l'interprétation, seul la valeur absolue importe et plus cette dernière est élevée, et plus l'utilisation de l'espace défini par cette variable est étroite. Il s'agit alors d'une mesure de la tolérance de l'espèce par rapport aux différentes conditions environnementales. Une espèce sera tolérante à une variable si le coefficient de cette variable sur l'axe de spécialisation est faible en valeur absolue et inversement une espèce sera très peu tolérante aux variations d'une variable ayant une spécialisation élévée en valeur absolue. Un telle variable sera un facteur de spécialisation pour l'espèce.

L'ENFA est le fruit des travaux de Hirzel *et al.* (2002) et représente une élégante réponse au problème de d'analyse exploratoire de la niche d'une espèce. Elle est utilisée en général

3. le biplot est une représentation simultanée du nuage des individus et celui des variables

pour réaliser des cartes d'habitats favorables. Mais cette méthode a été affinée par Basille *et al.* (2008) ; Calenge et Basille (2008), qui l'utilisent pour des analyses exploratoires des niches. Néanmoins, elle repose toujours sur deux principes clés : celui de marginalité et de spécialisation. Cependant, il existe une autre vision de la niche écologique basée sur un seule critère : la *distance de Mahalanobis* (Mahalanobis, 1936) et qui conduit à la seconde méthode factorielle : la MADIFA.

3.1.2 Mahalanobis Distance Factor Analysis : MADIFA

L'analyse factorielle des distances de Mahalanobis (MADIFA) est une méthode factorielle basée sur une décomposition en axes principaux de la distance de Mahalanobis entre les points disponibles et l'optimum de l'espèce dans l'espace écologique (point moyen). Afin de comprendre l'intuition qui motive cette décomposition factorielle, il semble nécessaire d'expliquer le rôle de la distance Mahalanobis dans les modèles de distribution d'espèce.

Distance de Mahalanobis et niche écologique

En statistiques classiques, il s'agit d'une distance utilisée depuis plus d'un demi siècle (Mahalanobis, 1936), comme une généralisation de la distance euclidienne classique. Il faut attendre les travaux de Clark *et al.* (1993) pour voir son introduction dans les modèles de distribution d'espèces. De manière plus précise le $\mathcal{D}^2$ de Mahalanobis dans le cadre des modèles de distribution d'espèce est une mesure de *dissimilarité* entre tous les pixels qui constituent l'ensemble des variables environnementales et la moyenne de ces pixels où vit l'espèce (Clark *et al.*, 1993 ; Knick et Dyer, 1997). Si on part du principe que l'optimum de la niche de l'espèce est défini par sa moyenne, alors plus les conditions environnementales seront proches des conditions moyennes où vit l'espèce et plus le $\mathcal{D}^2$ sera faible et donc le pixel sur lequel cette distance sera calculée sera considéré favorable pour l'espèce et inversement. Cependant, l'utilisation de la moyenne comme optimum de la niche a été critiquée par plusieurs auteurs (Rotenberry *et al.*, 2002, 2006 ; Browning *et al.*, 2005) comme n'étant pas une hypothèse réaliste. Il est alors possible dans ces conditions, qu'une distance de Mahalanobis élevée ne soit plus caractéristique d'une zone défavorable, mais plutôt d'un espace favorable à l'espèce. Donc en utilisant cette distance, on s'expose à un risque de biais dans l'estimation globale des zones d'habitat favorable. Partant de ce constat, Rotenberry *et al.* (2002, 2006) ; Dunn et Duncan (2000) ; Browning *et al.* (2005) proposent une variante de la méthodologie originale de Clark *et al.* (1993). Cette modification est basée sur l'idée qu'il ne faut plus utiliser la moyenne de la niche comme optimum, et qu'il faut dès lors chercher un critère plus réaliste. Leurs analyses les mèneront finalement à une partition linéaire de la métrique de Mahalanobis.

Le principe écologique à la base de leur méthodologie est de remplacer la moyenne de la niche par un ensemble de *conditions minimales qui permettent à l'espèce de survivre*. Cet ensemble de conditions minimales serait selon les auteurs plus proches des besoins de l'espèce et donc plus robuste dans l'analyse que la moyenne. L'intuition biologique est que les variables environnementales qui sont très peu variables dans les zones où l'espèce se trouve sont celles qui composent cet ensemble de prérequis minimaux dont l'espèce a besoin pour survivre. En effet, ils s'agit d'une *base minimale commune* à toutes zones occupées par l'espèce. De manière analytique, Rotenberry *et al.* (2002, 2006) procèdent en effectuant une ACP de la niche (l'espace utilisé) de l'espèce et proposent de ne garder pour leurs analyses que les derniers axes de l'ACP, qui par construction, sont caractérisés par une *très faible variance*. Ces axes de faibles variabilités peuvent dès lors être utilisés comme fondation pour cet ensemble minimal de variables, car elles sont caractéristiques de groupes de variables qui varient le moins possibles entre les zones occupées. Outre cette méthode pour calculer les conditions minimales requises par l'espèce, ,ils démontrent aussi que cette ACP est une manière de réaliser une partition linéaire de la distance de Mahalanobis. Donc ils reconstituent une distance de Mahalanobis basée sur les derniers axes de l'ACP de la niche, et à l'instar de la métrique classique de Mahanolobis, ils utilisent cette distance comme un indicateur d'habitats favorables.

On garde les mêmes notations matricielles que celles utilisées pour décrire l'ENFA. Soit Σ la matrice de corrélation de X sur l'espace utilisé ($\Sigma = X^\top D_p X$). Par définition, la distance de Mahalanobis entre un pixel x_i et la niche moyenne μ :

$$\mathcal{D}_i^2 = (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu)$$

Si on note $X_{i.}$ la ligne i de la matrice centrée X, on a :

$$\mathcal{D}_i^2 = X_{i.}^\top \Sigma^{-1} X_{i.}$$

On remarque que dans l'équation , il faut inverser la matrice Σ. Une façon de résoudre ce problème d'inversion est d'utiliser une diagonalisation de cette matrice [4], or les vecteurs propres issuent de cette diagonalisation correspondent au axes de l'ACP du tableau X en utilisant comme poids la matrice D_p : il s'agit alors d'une ACP de la niche.
Soit v_i un vecteur propre associé à la valeurs propre λ_i, on peut montrer que :

$$\mathcal{D}_i^2 = \sum_{j=1}^{p} \left(\frac{X_{i.} v_j}{\sqrt{\lambda_j}}\right)^2 = \sum_{j=1}^{p} b_{ij}^2$$

4. qui est symétrique donc diagonalisable dans une base orthonormée

Formellement, l'approche de Rotenberry *et al.* (2002, 2006), consiste alors à ne garder que les derniers vecteurs propres associés de cette décomposition (ici r axes), ceux de variances minimales.

$$\hat{\mathcal{D}}_i^2 = \sum_{j=p-r}^{p} b_{ij}^2$$

Cette décomposition linéaire du $\mathcal{D}^2$ et sa reconstitution est une amélioration de la méthodologie originale de Clark *et al.* (1993). Cette approche a été utilisée très longtemps, principalement dans les études de conservation de grands vertébrés. Néanmoins, malgré la base mathématique solide et l'intuition biologique derrière ce modèle, il existe un problème majeur qui a échappé aux auteurs, et qui n'a donc pas été pris en compte dans leurs analyses : *la prise en compte de l'espace disponible*.
Ce constat est le point de départ des travaux de Calenge *et al.* (2008). Leurs analyses reposent sur celles de Rotenberry *et al.* (2002, 2006). Ils conservent le concept de base qui est la recherche de *conditions minimales nécessaire à l'espèce pour survivre*.

Ils proposent alors un nouvel algorithme qui permet d'améliorer l'utilisation du $\mathcal{D}^2$ dans les modèles de distribution d'espèce. Cette nouvelle méthode permet une décomposition de cette métrique tout en tenant compte de l'espace écologique disponible. Cette nouvelle méthode est l'*Analyse Factorielle des Distances de Mahalanobis* (MADIFA).

Une amélioration de la méthode : MADIFA

La MADIFA permet de capter les directions dans l'espace écologique sur lesquelles la niche est la plus étroite par rapport à l'environnement disponible.

Par rapport à l'ENFA qui nous permet de quantifier la position (marginalité) et la forme de la niche (spécialisation), la MADIFA combine ces deux indicateurs en un seul, qui est une mesure globale de la restriction de la niche. Cet aspect fait que ces deux méthodes sont complémentaires, et Calenge *et al.* (2008) proposent d'utiliser l'ENFA afin de détecter parmi les axes de la MADIFA ceux qui compte le plus pour la marginalité ou la spécialisation.

Les deux premières étapes de la MADIFA sont les mêmes que celles de Rotenberry *et al.* (2002, 2006).

La MADIFA se démarque de leur méthode originale par calcul d'une seconde ACP sur la matrice B (dont les élements sont les b_{ij}). Cette seconde ACP permet de trouver des combinaisons linéaires des variables environnementales caractérisant une direction dans l'espace écologique où la niche est la plus étroite possible par rapport à l'espace disponible. Cette étroitesse est alors caractéristique de la base minimale commune à l'espèce.

Formellement, l'ACP de B se fait par diagonalisation de la matrice $G = B^\top DB$. Cette diagonalisation nous permet d'obtenir P vecteurs propres v_k. Soit V la matrice dont les v_k sont des vecteurs colonnes, les valeurs propres θ_k associées sont stockées dans la matrice diagonale $\mathcal{B}$. On note A la matrice de passage qui permet de réaliser la première ACP (diagonalisation de la matrice Σ) et soit δ la matrice diagonale qui contient les valeurs propres de cette diagonalisation. Donc on a :

$$G = V\mathcal{B}V^\top$$

Si on pose :

$$C = A\delta^{-\frac{1}{2}}V$$

Alors l'indicateur qui sera utilisé pour reconstruire une distance moyenne de Mahalanobis pour chaque pixels est :

$$L = XC$$

La matrice L contient les scores l_{ij}, la valeur qui est maximisée sur les premiers axes de la MADIFA est :

$$\theta_j = \frac{\sum_i^n \frac{1}{n}(l_{ij} - \bar{l}_j^u)^2}{\sum_i^n u_i(l_{ij} - \bar{l}_j^u)^2}$$

l_{ij} est le score du pixel i sur le j^{me} axe de la MADIFA. La valeur $\bar{l}_j^u$ représente la moyenne sur le j^{me} axe des scores des pixels de l'espace utilisé par l'espèce, u_i est l'élément i du vecteur u d'utilisation de l'espace.
Le dénominateur de θ_j est la variance de la niche sur les axes de la MADIFA. Mais le numérateur n'est pas une variance mais le carré moyen de la *déviation entre les points disponibles et la moyenne des scores de points utilisés*. Donc la MADIFA permet de detecter les directions selon lesquelles la niche est la plus étroite comparativement à l'espace disponible. Ces directions sont celles qui permettent d'obtenir l'*ensemble minimal de variables dont a besoin l'espèce pour survivre*.

En centrant la niche, la niche moyenne se trouve à l'origine des différents plans factoriels de l'ACP de B, donc :

$$\hat{l}_{ij}^u = 0$$

De plus, en réduisant les axes de la seconde ACP (matrice B), la variance de la niche sur les axes de la MADIFA vaut 1, donc :

$$\sum_{i}^{N} u_i (l_{ij} - \hat{l}_j^u)^2 = 1$$

Donc finalement en remplaçant, on obtient :

$$\theta_j = l_{ij}^2$$

Et Calenge *et al.* (2008) montrent ce résultat important :

$$\mathcal{D}^2 = \sum_{j} l_{ij}^2$$

Donc chaque valeur propre est associée à un axe de la MADIFA et explique une portion du carré de la distance de Mahalanobis moyenne. Il est dès lors possible alors de partitionner cette distance en utilisant les axes de la MADIFA.

Le sens biologique des axes de la MADIFA s'obtient en général en analysant la corrélation avec les variables environnementales qui y sont projetées et en utilisant les résultats de l'ENFA.

3.2 Applications et résultats

Les différentes variables choisies pour l'ENFA figurent dans le tableau 3.1. Il s'agit de l'amplitude et de la phase de la première harmonique de la décomposition de Fourier des images MODIS, de la moyenne, du maximum et de minimum des variables. Pour les variables dont la corrélation (sur données de présence) était supérieure au seuil de 0.95, nous avons choisi de garder celle dont l'explication biologique est plus simple. Cet ainsi que les moyennes qui sont toutes extrêmement corrélée au maximum ou minimum selon l'indicateur ont été abandonnée. Nous disposons aussi de 91 données de présence à travers notre zone d'étude. Les méthodes que nous utilisons sont purement descriptives donc nous utiliserons toutes [5] les observations. Les méthodes présentées ci-dessus sont des approches exploratoires et donc l'inférence classique n'est pas possible. Donc pour s'assurer qu'il y a effectivement selection d'habitat significative par l'espèce (au moins dans une directon), nous utiliserons une approche non paramétrique basée sur le bootstrap (Davison et Hinkley, 1997). On utilisera en particulier les tests de Monte-Carlo. Il s'agit de faire un tirage sans remise des pixels utilisés (points de présence), donc de simuler une niche aléatoire pour l'espèce sur la zone d'étude et de calculer à chaque simulation une statistique (valeurs propres, marginalité, etc.). On répète ce processus plusieurs fois (entre 500 et 1000), on a

5. pas d'utilisation de jeu de donnée de calibration et de jeu de donnée de test

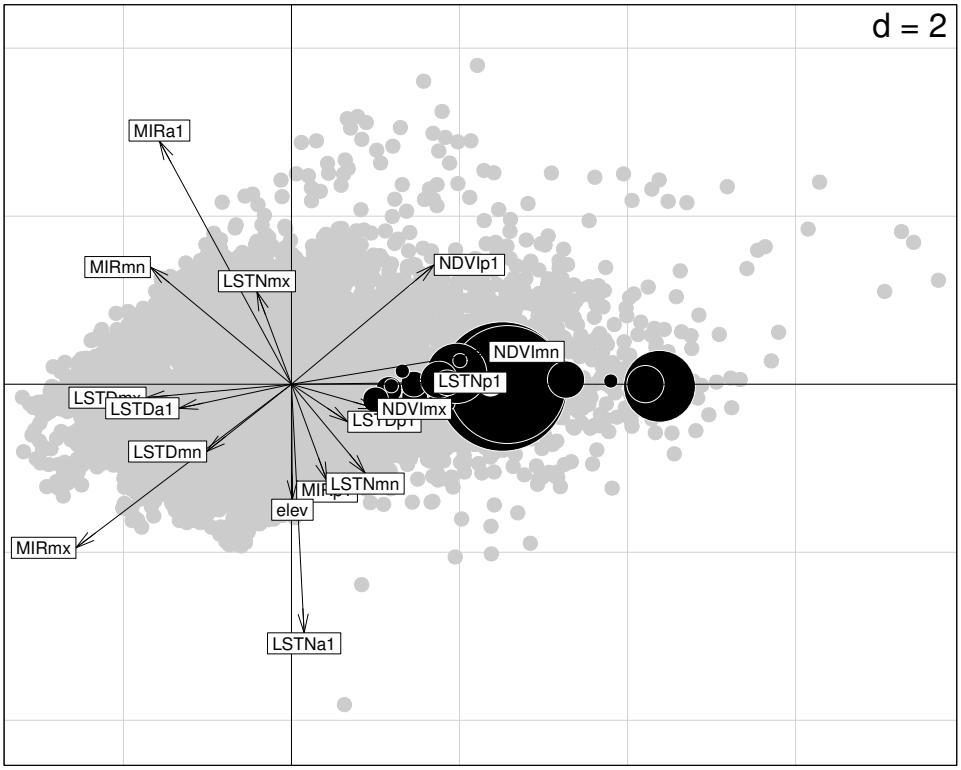
ainsi une distribution bootstrap sous l'hypothèse de choix aléatoire de l'habitat par l'espèce. Notre statistique est ensuite comparée à cette distribution. Ce test non paramétrique simple, permet alors d'avoir une p-value qui est signifactive en cas de choix d'habitat.

Tableau 3.1 : Définition des variables

espace		disponible		utilisé	
Nom	Description	moyenne	amplitude[1]	moyenne	amplitude
elev	élévation	29.301	130.000	28.026	74.000
LSTDa1	LST day première amplitude	3.453	6.300	2.480	3.780
LSTNa1	LST night première amplitude	2.748	2.640	1.520	2.839
MIRa1	MIR première amplitude	0.091	0.148	0.068	0.075
LSTDmn	LST day minimum	28.916	15.860	28.150	5.080
LSTNmn	LST night minimum	15.193	6.840	15.541	3.660
LSTDmx	LST day maximum	40.1583	19.240	37.392	10.540
LSTNmx	LST night maximum	22.287	4.060	22.248	1.880
MIRmn	MIR minimum	0.238	0.429	0.195	0.138
MIRmx	MIR maximum	0.446	0.546	0.364	0.162
NDVImn	NDVI minimum	0.178	0.488	0.209	0.149
NDVImx	NDVI maximum	0.445	0.783	0.489	0.400
LSTDp1[2]	LST day phase 1	3.000	11.000	3.000	9.000
LSTNp1	LST night phase 1	9.000	3.000	9.000	2.000
MIRp1	MIR phase 1	4.000	4.000	4.000	1.000
NDVIp1	NDVI phase 1	10.000	8.000	10.000	2.000

source : TALA, calculs :auteur

En moyenne les variables liées à la végétation (NDVImn) ont des valeurs plus élevées dans les zone occupées que dans celles disponibles, on observe l'inverse pour les indicateurs de réflectance infrarouge (MIR) et de température de jour (LSTD). En moyenne *G. p. gambiensis* ne semble pas affecté par la variable élévation, mais au niveau de l'amplitude (maximum - minimum), on note une variabilité plus faible. On remarque aussi que l'espèce préfère des gammes de variation réduites pour les indicateurs de végétation.



Graphique 3.3 : Biplot du premier plan factoriel

La première valeur propre est de 51.23, soit 41% (explication), cette valeur propre est significative avec une p-value inférieur à 0.001, donc on peut conclure à une sélection d'habitat significative sur le premier axe de l'ENFA. De plus avec une marginalité de 5.61 (p-value inférieure à 0.001) l'habitat de *G. p. gambiensis* diffère fortement des conditions moyennes disponible dans la zone des Niayes.

L'analyse du plan factoriel, permet de remarquer que les variables liées au MIR présentent une aussi forte marginalité qu'un critère de spécialisation (graphique 3.3) .La niche de *G. p. Gambiensis* est très sensible à une augmentation du rayonnement infrarouge.

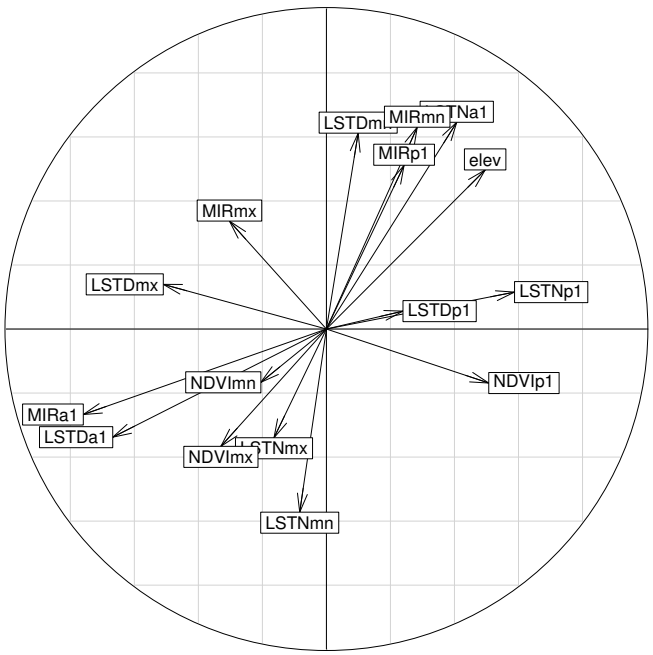
Sur l'axe de marginalité, On peut aussi noter que l'espèce préfère des zones où les températures maximales enregistrées (LSTDmx) sont beaucoup plus faibles que la moyenne disponible sur la zone des Niayes. Il faut aussi noter que les zones favorables sont fortement influencées par les mouvements saisonniers annuels de la température et de la végétation. Chaque année, pendant la saison des pluies, le NDVI est à son maximum pour certaines plantes (plantes annuelles herbacées) mais pas pour la végétation arborée où il semble être au maximum en novembre. L'activité chlorophyllienne est minimale en saison sèche et à ce moment de l'année, la recherche des refuges est un paramètre important pour les glossines riveraines qui vont chercher des zones humides arborées où elles s'abriteront. Le NDVImn est une variable critique dans le choix de l'habitat, ce qui se matérialise par une marginalité élevée (0.418). Parallèlement on observe que cela est corrélé au LSTNp1, ce qui correspond

à un effet tampon de la végétation arborée qui empêche le refroidissement nocturne en bloquant le rayonnement infrarouge. On observe également une corrélation négative avec le MIRmx, caractéristique d'un rayonnement maximal au niveau des sols nus.
Sur l'axe de spécialisation on constate que les glossines sont très sensibles à des variations de l'amplitude annuelle du MIR et du LSTN. Cette faible tolérance s'explique probablement par le rôle tampon de la végétation.

Tableau 3.2 : Coordonnée des variables sur les différents axes de l'ENFA

	Mar (41%)	Spe1 (19%)	Spe2 (9%)
MIRmx	-0.456	-0.346	-0.487
NDVImn	0.418	0.069	-0.050
LSTDmx	-0.308	-0.028	0.075
NDVIp1	0.302	0.253	0.259
LSTNp1	0.301	0.003	0.015
MIRmn	-0.299	0.248	0.693
MIRa1	-0.280	0.515	0.371
LSTDa1	-0.238	-0.051	-0.053
NDVImx	0.182	-0.052	0.188
LSTDmn	-0.180	-0.142	-0.094
LSTNmn	0.156	-0.188	-0.069
LSTDp1	0.119	-0.079	0.036
MIRp1	0.075	-0.205	-0.068
LSTNmx	-0.073	0.196	-0.000
LSTNa1	0.027	-0.526	-0.087
elev	0.002	-0.243	-0.063

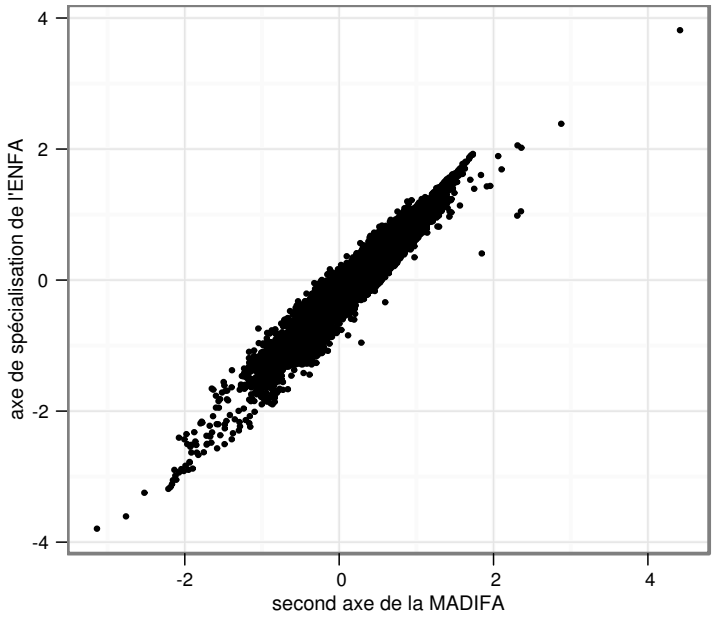
La première valeur propre de la MADIFA, vaut 157.38 (p-value = 0.002), donc il y a sélection d'habitat significative par l'espèce.



Graphique 3.4 : Corrélation entre les variables environnementales et les axes de la MADIFA

Le premier axe de la MADIFA est caractérisé une opposition entre deux groupes de variables. D'un côté nous avons des variables de phases liées à la saisonnalité (variable de marginalité) et d'un autre côté les amplitudes qui caractérisent une très forte spécialisation (ENFA). Donc le premier axe est défini par des variables qui comptent autant pour la marginalité que pour la spécialisation.

Quant au second axe de la MADIFA, il est très corrélé ($r_{pearson}$ = 0.96, p-value < 0.001, graphique 3.5) à l'axe de spécialisation de l'ENFA. Donc la direction que detecte la MADIFA à travers cet axe est la même que l'ENFA a décelée sur son premier axe de spécialisation : Les variables pour lesquelles l'espèce est la moins tolérante sont celles qui définissent alors le second axe de la MADIFA.



Graphique 3.5 : Relation entre le second axe de la MADIFA et le premier axe de spécialisation de l'ENFA

Donc la MADIFA, permet de renforcer les résultats obtenus avec l'ENFA. Nous avons calculé la norme des variables (longueur des flèche dans le plan) sur les deux plans factoriels de l'ENFA (second plan en annexe A). Les variables les plus importantes sont celles qui comptent le plus pour la marginalité et la spécialisation. Nous avons alors choisi celles dont cette norme est supérieure à la norme médianne sur les deux plans factorielles (valeurs propres et second plan en annexe A). Ces variables seront utilisées dans la suite pour prédire les zones d'habitats favorables.

Finalement, l'application de ces méthodes factorielles nous a permis de mettre en évidence :

- Une sélection significative de l'habitat par l'espèce
- Un habitat favorable caractérisé par une végétation arborescente
- Que les zones où les indicateurs de réflectance infrarouge et de température sont élevés sont peu favorables

Chapitre 4

Prédiction de la niche potentielle

Plusieurs méthodes prédictives sont souvent utilisées pour modéliser la niche de l'espèce et prédire sa distribution géographique. Ces méthodes sont d'autant plus efficaces et utiles dans le cadre d'un modèle conceptuel théorique bien établi (Guisan et Thuiller, 2005).

Le choix des différents modèles utilisés a un grand impact sur le résultat final. Cette différence est d'autant plus grande si les méthodes n'utilisent pas les mêmes type de donnée de terrain. Parmi les modèles de distributions d'espèces (chapitre 1) nous avons distingué 4 groupes selon le type de données utilisés pour l'algorithme. Dans ce chapitre, nous modéliserons la niche potentielle [1] de *G. p. gambiensis* en utilisant trois de ces modèles :

- Un modèle de présence-seule
- Un modèle de présence-background
- Un modèle de présence-absence

Les modèles de présence-pseudoabsence n'ont pas été considérés car ils sont assez proches des modèles de présence-background en théorie et en pratique les algorithmes utilisés sont souvent les mêmes que ceux de présence-absence.

L'analyse et la comparaison de ces différents modèles nous permettra d'avoir une vision plus complète de la distribution potentielle de l'espèce d'étude. De plus, en utilisant des modèles différents on s'affranchit de l'effet intrinsèque lié au choix du modèle utilisé sur les différents résultats.

4.1 Méthodologie

Les trois modèles qui seront présentés, sont différents par le type de donnée utilisé mais aussi par les algorithmes qu'ils utilisent. Le premier modèle est basé sur la métrique de

1. sa projection dans l'espace géographique

Mahalanobis et est un modèle de présence-seule, le second est un modèle d'apprentissage statistique basé sur la notion d'entropie : MaxEnt et finalement on utilisera la méthode des forêts aléatoire comme modèle de présence-absence.

4.1.1 un modèle de présence-seule : la Distance de Mahalanobis

La distance de Mahalanobis fait partie de méthodes utilisant juste la connaissance des données d'occurrences Dunn et Duncan (2000). Le principe est le même que celui qui a été présenté au chapitre 3, mais dans cette section, nous l'utiliserons pour avoir une distribution de probabilité sur l'espace d'étude (tous les pixels).

La distance de Mahalanobis entre un point y de l'espace écologique et la niche moyenne μ vaut :

$$D^2(y) = (y - \mu)^\top \Sigma^{-1}(y - \mu)$$

Σ^{-1} est la matrice de corrélation des variables décrites sur la niche. On peut montrer que $\mathcal{D}^2$ suit approximativement un $\chi^2_{(p\ ddl)}$ (p le nombre de variable) sous l'hypothèse de multinormalité des variables environnementales. Ce résultat important permet alors d'obtenir une carte de probabilité, de p-values. Ces probabilités sont analogues aux probabilités à posteriori calculées en analyse discriminante linéaire : il s'agit alors d'indicateurs d'habitat favorable. Nous avons

$$p = 1 - prob(\chi^2_P)$$

Cette probabilité est calculée sur chaque pixel, et on obtient ainsi une carte de probabilité de présence sur la zone d'étude.

4.1.2 Un modèle de présence-background : MaxEnt

Maximum entropy (MaxEnt) (Phillips *et al.*, 2006) est un modèle de distribution d'espèce qui n'utilise que les données de présence. Elle fait partie de la famille des méthodes d'apprentissage statistique au sens de Hastie *et al.* (2009). Elle est basée sur le principe d'entropie maximum qui établit que la meilleure approximation d'une distribution inconnue est celle qui maximise l'entropie (la plus proche de la distribution uniforme) sous certaines contraintes.
Elle se base sur la minimisation de l'entropie entre la fonction de densité des données de présence et celles de l'environnement. Un des avantages de l'algorithme est la possibilité d'utiliser des modèles très complexes qui prennent en compte une large gamme d'interactions entre les différentes variables explicatives disponibles. Nous reprendrons les

notations de Elith *et al.* (2011) et l'explication de la méthode va se faire dans l'espace écologique. Pour une explication de la méthode dans l'espace géographique, voir Phillips *et al.* (2006).

Soit L un sous ensemble de l'espace écologique qui représente la zone d'étude, l'évènement $y = 1$ dénote une présence de l'espèce et $y = 0$ son absence. On note z le sous échantillon de L qui est constitué par les variables environnementales et climatiques disponibles. Soit $f_1(z)$ la fonction de densité multivariée des variables environnementales dans la zone où se trouve l'espèce et on note $f(z)$ la fonction de densité de ces variables environnementales sur L (l'espace disponible). On a donc :

$$f_1(z) = Pr(z \mid y = 1)$$

Par rapport à f_1, f est une densité marginale, inconditionnelle. La fonction de densité f représente alors la distribution de l'espèce dans le cas où cette dernière serait complètement insensible aux variables environnementales. Sans la présence de variables environnementales, il n y a aucune raison de penser que l'espèce choisirait un zone plutôt qu'une autre, donc sa probabilité d'occurrence serait *uniforme* sur L. Le but de la méthode est de modéliser la probabilité de présence, conditionnellement à l'environnement disponible :

$$Pr(y = 1 \mid z)$$

En utilisant la règle de Bayes on obtient :

$$Pr(y = 1 \mid z) = \frac{Pr(z \mid y = 1)Pr(y = 1)}{Pr(z)} = \frac{f_1(z)Pr(y = 1)}{f(z)}$$

Avec les données de présence-background, on a alors la possibilité d'estimer les densités f_1 et f mais pas le terme $Pr(y = 1)$ qui représente la prévalence de l'espèce dans la zone d'étude. Sans l'utilisation des données d'absences, cette prévalence n'est pas identifiable (Ward *et al.*, 2009). Donc la modélisation de la quantité d'intérêt ($Pr(y = 1 \mid z)$) est possible mais à une constante près, la prévalence. MaxEnt estime dans un premier temps le rapport f_1/f pour obtenir cette probabilité. Ce ratio est calculé en estimant la distribution de $f_1(z)$ la plus proche possible de $f(z)$ mais sous la contrainte que la valeur des variables dans les zones occupées soit le *reflet de cette densité*. En moyenne, les réalisations de la densité f_1 ne doivent pas être très différentes des valeurs moyennes des variables sur l'espace utilisé. Le but est alors de comparer ces deux densités en minimisant l'entropie relative entre ces deux quantités.

Phillips *et al.* (2006) montre que minimiser l'entropie relative entre ces deux densités, conduit à un modèle de la famille exponentielle (distribution de Gibbs) et on a :

$$f_1(z) = f(z)e^{\eta(z)}$$

avec $\eta(z) = \alpha + \beta.h(z)$, où $h(z)$ est le vecteur des variables explicatives, et α une constante qui permet de normaliser f_1 (pour que son intégrale fasse 1).

Donc le but de l'algorithme est alors de modéliser $e^{\eta(z)}$ qui est une estimation du ratio des deux densités. On retrouve alors un modèle log-linéaire qui dépend que des présences et du background. Le programme d'optimisation qui permet d'obtenir les coefficients β et donc qui minimise l'entropie relative est :

$$\max_{\beta,\alpha} \frac{1}{m} ln(f(z_i)e^{\eta(z)}) - \sum_j \beta_j |\lambda_j|$$

sous contrainte que

$$\int_L f(z)e^{\eta(z)}dz = 1$$

Les coefficients λ sont des coefficients de régularisation, ils permettent de pénaliser l'entropie relative et de faire un compromis entre complexité (absence de biais) et parcimonie (variance minimale des estimations et prédiction) pour donner une bonne valeur prédictive au modèles. Des modèles trop complexes font souvent du sur-ajustement (overfitting) et ne permettent pas de généraliser les résultats. Donc le choix de ces coefficients permet de faire un bon compromis entre complexité et généralisation du modèle.

4.1.3 Un modèle de présence-absence : les forêt aléatoires

La technique des forêts aléatoires (random forest) est une méthode d'apprentissage statistique (Hastie *et al.*, 2009) introduite par Breiman (2001). Elle fait partie des méthodes d'ensembles et est basée sur la notion de bagging (Boostrap AGGregatING) (Breiman, 1996) et d'arbres de décision.

Le bagging

Dans la littérature des méthodes d'apprentissage statistique, un des problèmes majeurs est de résoudre le compromis entre biais et variance (bias-variance tradeoff). En effet, les méthodes estimations favorisent l'un ou l'autre selon les objectifs visés par une étude. Le bagging est une technique qui permet de réduire la variance d'un estimateur, sans pour autant beaucoup augmenter le biais de cet estimateur(Breiman, 1996). Il s'agit d'une méthode qui permet d'obtenir de bons résultats pour les techniques d'estimations connues pour donner des estimateurs de faibles biais et de grandes variances, comme les arbres

de décision. Supposons que nous disposons d'un échantillon d'apprentissage pour calibrer notre modèle. On appellera ce modèle règle de base (e.g arbres de décision), cette règle permet de construire sur cet échantillon une estimation ou prédicteurs. Le principe du bagging est de tirer indépendamment plusieurs échantillons bootstrap et d'appliquer une règle de base sur chacun de ces échantillons. On obtient ainsi un ensemble de prédicteurs qu'on agrège par la suite pour avoir un prédicteur final plus performant. Pour une démonstration des propriétés du bagging, voir Breiman (1996).

Les arbres de décision : CART

Il s'agit de la règle de base généralement utilisée en bagging. Un algorithme d'arbres de décision populaire est Classification and Regression Trees (CART). CART désigne une méthode statistique introduite dans les années 80 par Breiman (1984) qui construit un prédicteur pour résoudre des problèmes de régression ou de classification. Le principe de CART est de partitionner récursivement l'espace des individus de façon dyadique avant ensuite de déterminer une partition optimale pour la prédiction. A chaque partition, l'espace est séparé en deux sous-parties. Cette partition est représentée par un arbre binaire, dont les nœuds sont les éléments (ici les deux sous-parties). Et on partitionne ainsi chaque sous partie, selon une règle de découpe précise. Cette règle est basée sur des critères à optimiser. Pour les problèmes de régression, on construit des nœuds de variance minimale et pour les problèmes de classification on cherche a obtenir des nœuds homogènes. On utilise une règle d'arrêt afin de développer l'arbre ainsi construit et les nœuds terminaux (non découpé) qui sont appelés feuilles. Cet arbre maximal, est ensuite élagué car l'arbre maximal possède une très grande variance et un biais faible. Le prédicteur final est alors une fonction constante par morceaux de sous arbre optimal (issue de l'élagage de l'arbre maximal).

Les forêts aléatoires

La méthode des forêts aléatoires est une modification du bagging qui permet de construire un grand ensemble d'arbres de décision (variante de CART) non corrélés entre eux dans le but de les agréger ensuite. Dans la construction de ces arbres, pour découper chaque nœud, on tire aléatoirement un nombre m de variables parmi celles disponibles, et on cherche la meilleure sous-partition optimale uniquement suivant ces m variables. Cette étape de tirage d'un nombre donné de variables est la principale différence entre les forêts aléatoires et le bagging.

En pratique, les forêts aléatoires améliorent les performances du bagging (Breiman, 2001). L'explication heuristique de ce gain de performance s'explique par l'aléa supplémentaire qu'entraîne le tirage des variables dans la construction des arbres. Cette dose d'« aléa » rend

alors chaque arbre différent des autres, sans pour autant dégrader leurs performances individuelles. Il s'en suit que le prédicteur final agrégé est meilleur que celui obtenu de manière classique par un bagging simple. Cependant, si une trop grande perturbation est introduite pour construire les arbres, alors les prédicteurs individuels seront trop similaires et le prédicteur final n'apportera aucune amélioration. La réussite de l'algorithme dépend grandement du choix de m, qui doit permettre d'introduire une bonne dose d'aléa. La méthode des forêts aléatoires est beaucoup plus complexe et pour une revue complète de cette la référence, est Breiman (2001). Pour une revue récente en français, voir Genuer (2010).

4.1.4 Évaluation et comparaison des différents modèles

Évaluer et comparer le pouvoir prédictif de différents modèles est une étape critique dans le processus de modélisation de la niche d'une espèce. Nous suivrons l'approche Elith *et al.* (2006) qui évaluent et comparent des modèles de présence-seule, présence-background et présence-absence en utilisant des données de validations de présences et d'absences.

Nous avons construit trois modèles différents, pour prédire les zones potentiellement favorables à *G. p. gambiensis*. Ces modèles ont été choisis pour leurs approches différentes, cette différence s'exprime par les inputs nécessaires à la mise en place de chaque modèle. Néanmoins, malgré cette différence, la comparaison de ces modèles est possible en utilisant des critères communs et objectifs aux trois modèles. Nous utiliserons cinq critères de qualité de modèle :

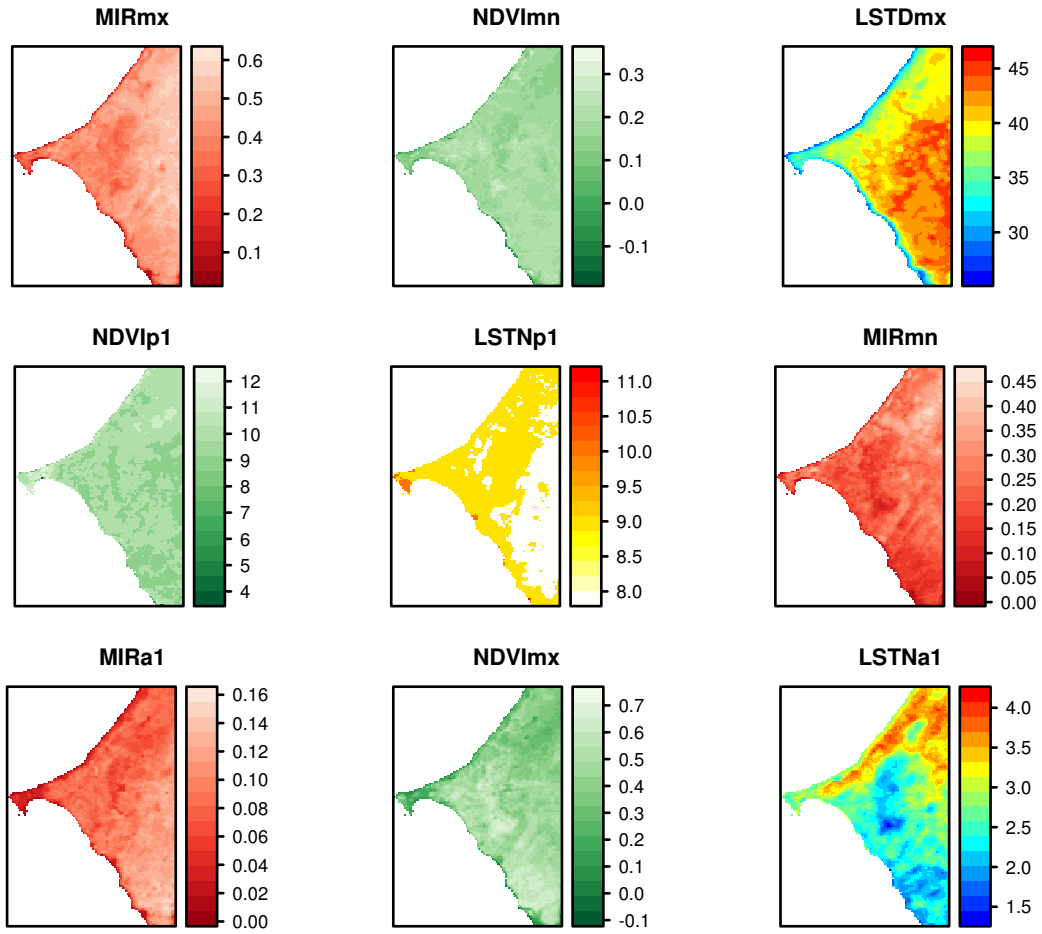
- L'aire sous la courbe ROC (Receiver Operator Curve), qu'on appelle généralement AUC (Area Under Curve) (DeLong *et al.*, 1988). Cet indicateur varie entre 0 et 1, un score de 1 indique une discrimination parfaite, un score de 0.5 caractérise un pouvoir discriminant aussi bon que la chance (aléatoire), et une AUC inférieure à 0.5 indique une discrimination moins bonne que la chance. C'est une statistique qui ne dépend pas d'un seuil, elle est par ailleurs proche de la statistique de Mann-Whitney (le U de Mann-Whitney), donc elle est à ce titre une statistique du rang.
- Le Kappa (Cohen *et al.*, 1960) est un indicateur qui permet de chiffrer l'accord entre les observations. Elle estime un taux de concordance entre les observations et leurs prédiction, en tenant compte des erreurs d'omission et de commission. Plus elle est élevé et plus le modèle est de bonne qualité ;
- le pourcentage de bien classé (PCC) est un indicateur couramment utilisé en classification. Sans l'apport de statistique sur la sensibilité et la spécificité son utilité est réduite ;
- la spécificité (Sp) est la probabilité d'avoir un résultat négatif sachant que l'individu est négatif (probabilité de vrai négatif) ;

– la sensibilité (Se) est probabilité d'avoir un résultat positif sachant que l'individu est positif (probabilité de vrai positif). On note par ailleurs que la courbe ROC est aussi la représentation graphique de la fonction de la probabilité de vrai positif (Se) en fonction celle de faux positif (1 - Sp).

Ces critères sont très souvent utilisé dans la comparaison des modèles de distribution d'espèce (Elith *et al.*, 2006 ; Hirzel *et al.*, 2006) et permettent alors de classer les modèles selon leur pouvoir prédictif.

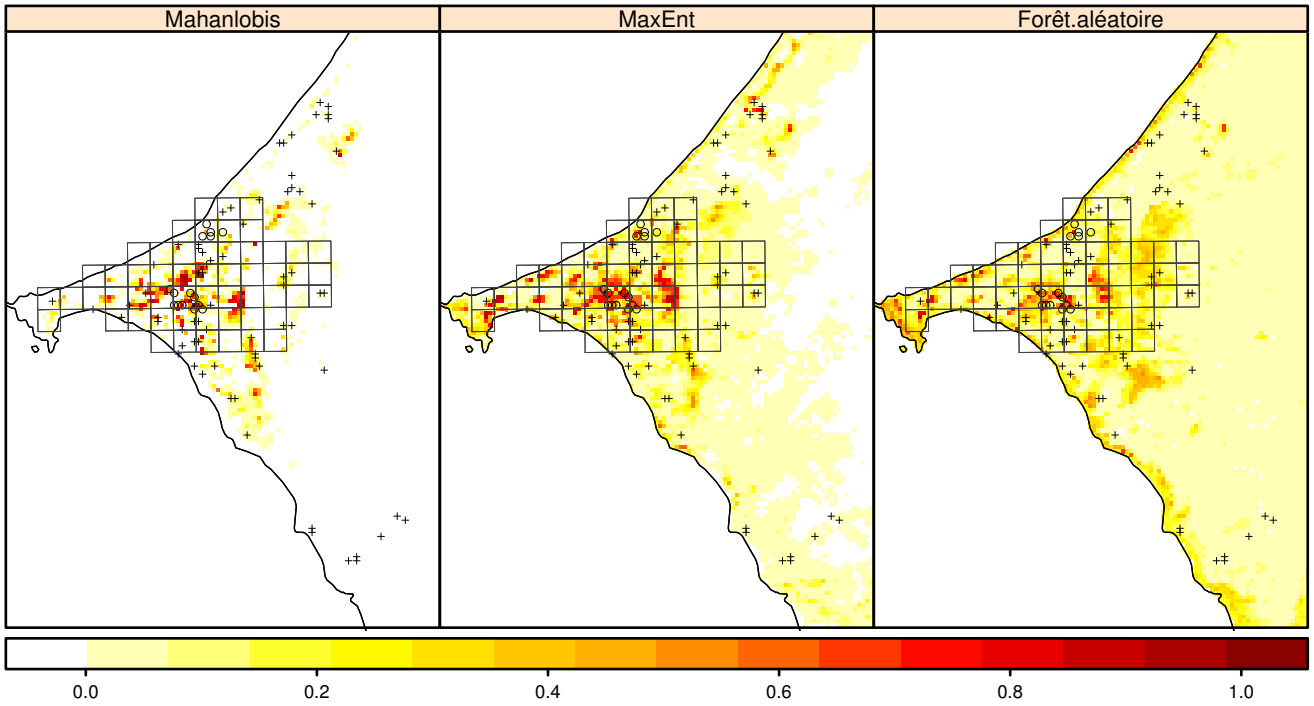
4.2 Applications et résultats

Contrairement aux méthodes exploratoires présentées en chapitre 3, les modèles que nous utilisons ne prennent pas en compte le fait que dans un même pixel on puisse avoir plus de deux points. Donc nous avons gardé pour ces algorithmes une seule présence par pixel. Ceci ramène notre nombre de présence à 38 (contre 91 avant) et le nombre de points d'absence à 218 (contre 369 avant). Le jeu de données de présence et d'absence a été partagé en un jeu de calibration ($\frac{2}{3}$) et un jeu de validation ($\frac{1}{3}$). Toutes les premières analyses de qualité du modèle seront alors faites sur ce jeu de validation. Cependant, en plus de ce jeu de validation, des données externes seront aussi utilisées. En effet, des relevés phytosociologiques ont été effectués à travers la zone d'étude indépendemment de l'échantillonnage (Bouyer *et al.*, 2010b). Ces données permettent d'avoir une classification de nos pixels en zones favorables (18 points en choisissant un point par pixel sinon 40) et non favorables (120 points en choisissant un point par pixel sinon 165). Il s'agit d'une classification à dire d'expert, de gîtes potentiels pour *G. p. gambiensis*. Ces informations seront utilisées pour mesurer le pouvoir prédictif des différents modèles et donnent un point de vue plus « qualitatif » sur ces modèles.



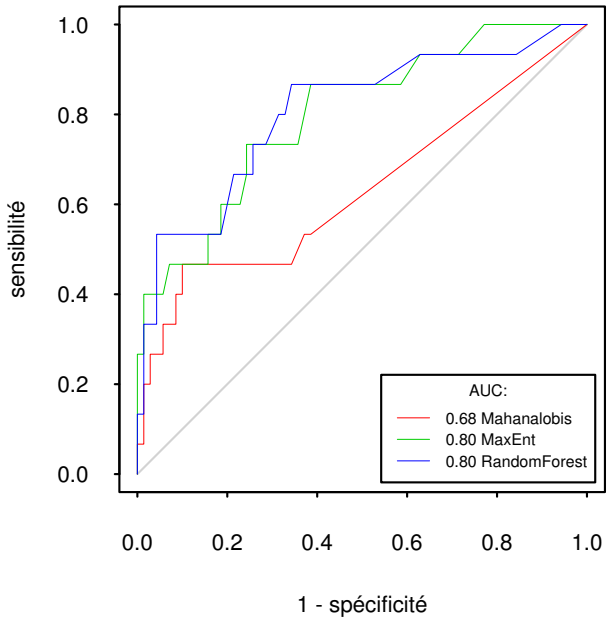
Graphique 4.1 : Prédicteurs utilisés pour les différents modèles.

Les variables environnementales utilisées (graphique 4.1) sont celles qui ont été choisies en utilisant l'ENFA et la MADIFA : on s'est assuré de garder l'ensemble de variables qui représentent les conditions minimales dont a besoin *G. p. gambiensis* pour sa survie. En gardant ces variables pour l'analyse, on a alors la possibilité d'avoir des modèles plus facilement interprétables, en relation avec l'écologie de l'espèce étudiée.



Graphique 4.2 : Probabilité d'occurence de *G. p. gambiensis*. La grille représente la zone de lutte a priori, les + sont des points d'absence et le o des points de présence (jeu de validation).

Pour les trois modèles, nous avons des probabilités de présence élevées au cœur de la zone de lutte, autour de Pout (graphique 4.2). La distance de Mahalanobis est beaucoup moins lisse dans son estimation (beaucoup de pixels de probabilité nulle). Les forêts aléatoires semblent indiquer que les côtes ont des probabilités d'occurrence assez élevées, alors que MaxEnt semble être compromis entre les deux premières estimations.



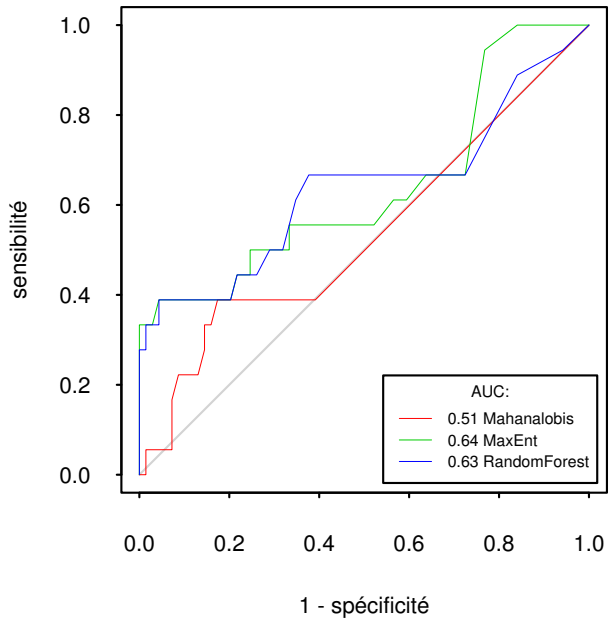
Graphique 4.3 : Courbe ROC. L'AUC de Mahalanobis est la plus faible, MaxEnt et RandomForest ont des résultats semblables.

L'AUC permet de classer les différents modèles. Le moins bon modèle est celui qui utilise la distance de Mahalanobis. On remarque que MaxEnt et les forêts aléatoires ont une AUC identique à 0.8. Ce qui est un indicateur de bon pouvoir prédictif de ces deux modèles.

Tableau 4.1 : Comparaison des différents modèles sur jeu de validation.

	Indicateurs				
Modèle	AUC	PCC (%)	Kappa	spécificité	sensibilité
Mahalanobis	0.680	82.350	0.301	0.928	0.333
MaxEnt	0.800	84.710	0.428	0.928	0.467
RandomForest	0.802	85.880	0.342	0.986	0.267

Les indicateurs de qualité du modèle présentés dans le tableau 4.1, ont été calculés en fixant le seuil de probabilité de présence à 0.5. Ce choix n'est pas optimal pour chaque modèle, mais il permet de les comparer plus facilement. Les indicateurs ainsi calculés sont en faveur des modèles MaxEnt et forêts aléatoires (RandomForest) qui ont des métriques semblables, sauf pour la sensibilité et la spécificité. Mahalanobis a une spécificité identique (0.928) à celle de MaxEnt et une sensibilité supérieure à celle obtenue par les forêts aléatoires. On note que le modèle qui a la plus grande sensibilité (voir annexe B pour une définition) est MaxEnt (0.467).
Nous avons réitéré la même analyse de qualité sur les données phytosociologiques.



Graphique 4.4 : Probabilité d'occurence de *G. p. gambiensis* sur relevés phytosociologiques.

Les AUC sont très faibles quand on utilise les données auxiliaires comme jeu de validation. Mahalanobis fait a peine mieux qu'une prédiction aléatoire.

Tableau 4.2 : Comparaison des différents modèles sur données auxiliaires.

	Indicateurs				
Modèle	AUC	PCC (%)	Kappa	spécificité	sensibilité
Mahalanobis	0.510	74.71	-0.022	0.928	0.055
MaxEnt	0.693	81.609	0.361	0.927	0.389
RandomForest	0.631	83.908	0.349	0.985	0.2778

On remarque une baisse au niveau de la qualité globale de tous les modèles. Cette baisse est d'autant plus prononcée pour le modèle basé sur la distance Mahalanobis. La sensibilité la plus élevée est obtenues par MaxEnt alors que comme pour la première validation, la méthode des forêts aléatoires semble se distinguer par une spécificité élevée.
L'utilisation de ce jeu de données permet de relativiser la qualité globale des différents modèles. La carte de distribution de probabilité avec les gîtes potentiels est disponible en annexe B.

Choix du seuil optimal

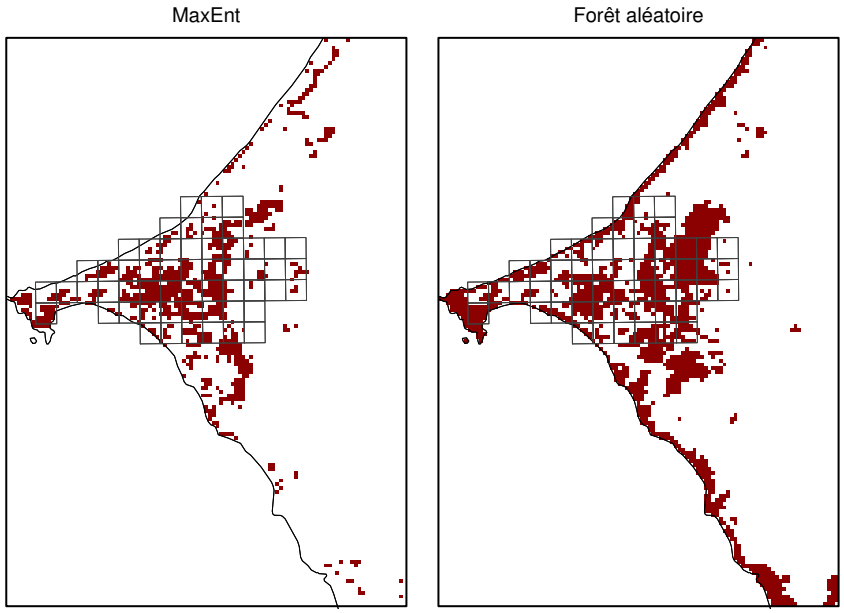
Les différents modèles nous ont permis d'obtenir des cartes de probabilité de présence sur tous les pixels de la zone d'étude. Ces cartes nous donnent des informations importantes sur les zones favorables à *G. p. gambiensis*. Cependant, pour délimiter la zone de lutte, il est nécessaire de choisir un *seuil* au dessus duquel le pixel de la zone sera considéré comme infesté par l'espèce. Une méthode simple et classique consiste à choisir un seuil fixe souvent choisi à 0.5. Pour Liu *et al.* (2005), ce choix de seuil est un « mauvais » choix, ils préconisent d'utiliser plusieurs critères de qualité du modèle pour faire ce choix. Suivant une approche similaire, nous allons choisir quatre critères différent, pour avoir une meilleure approximation du « bon » seuil :

- Le premier critère est basé sur la maximisation de la sommme de la spécificité et de la sensibilité. Ce choix de seuil correspond a un compromis entre ces deux quantités.
- Le second critère correspond a l'obtention d'une sensibilité égale à la spécificité. Il s'agit d'un seuil où la probabilité de détecter un pixel infesté est la même que celle d'identifier un pixel où l'espèce est absente comme non infesté.
- Le troisième critère est de minimiser la distance entre la courbe ROC et le point supérieur gauche du graphique de cette courbe (de coordonnées (0, 1)).
- Enfin, il est possible d'imposer une sensibilité donnée et de trouver le seuil qui permet de l'obtenir. Cette dernière méthode est très utile dans le cadre de notre étude où nous voulons minimiser le risque de ne pas détecter une population et donc maximiser la sensibilité.

Tableau 4.3 : Choix des seuils optimaux.

	Modèles		
Critère	Mahalanobis	MaxEnt	RandomForest
Max (Sens+Spec)	0.38	0.27	0.10
Sens=Spec	0.01	0.25	0.15
MinROCdist	0.38	0.27	0.14
ReqSens(= 0.75)	0.00	0.15	0.14

Nous choisirons donc le quatrième critère, qui n'est applicable que pour les derniers modèle, MaxEnt et les forêts aléatoires. Le choix de ce critère est motivé par l'objectif de l'étude : la délimitation d'une zone de lutte. En effet, le coût d'une évaluation trop optimiste de la zone favorable est moindre que l'inverse, ce qui alors peut entrainer un échec de l'éradication.
Nous avons donc choisi les seuils (un pour chaque modèle) permettant de fixer la sensibilité à 0.75 pour chaque modèle (tableau 4.3 , dernière ligne). Le choix de ce seuil n'est pas arbitraire et a été établi avec l'entomologiste principal du projet d'éradication.

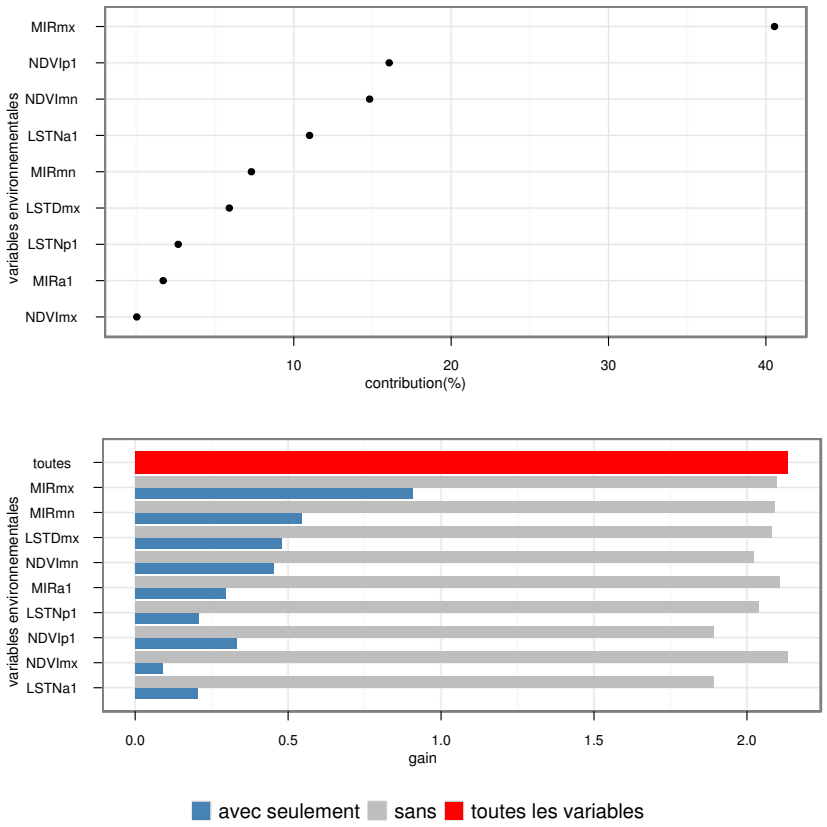


Graphique 4.5 : Zone d'habitats favorables avec choix de seuils permettant d'atteindre une sensibilité de 0.75. La grille représente la zone de lutte a priori.

L'habitat le long de la côte prédit favorable par la méthode des forêts aléatoires correspond à un artéfact lié à l'interface avec la mer, qui a également un effet tampon sur les paramètres thermiques du climat. Nous savons par ailleurs que cette côte n'est pas infestée (Bouyer, com. pers.). De plus, on note que cet artéfact ne semble pas affecter l'estimation faite avec MaxEnt. Ce dernier modèle apparaît alors comme celui qui réalise la meilleure estimation de la niche potentielle de l'espèce.

4.2.1 Influence des variables sur le modèle final (MaxEnt)

Le rôle que les variables environnementales joue dans l'estimation de la niche potentielle est très important. La contribution de chaque variable au modèle est calculée en utilisant un jacknife [2] sur l'espace des variables. L'algorithme derrière MaxEnt est itératif donc il est possible à chaque itération de calculer le gain qu'engendre l'omission ou la commission d'une variable.



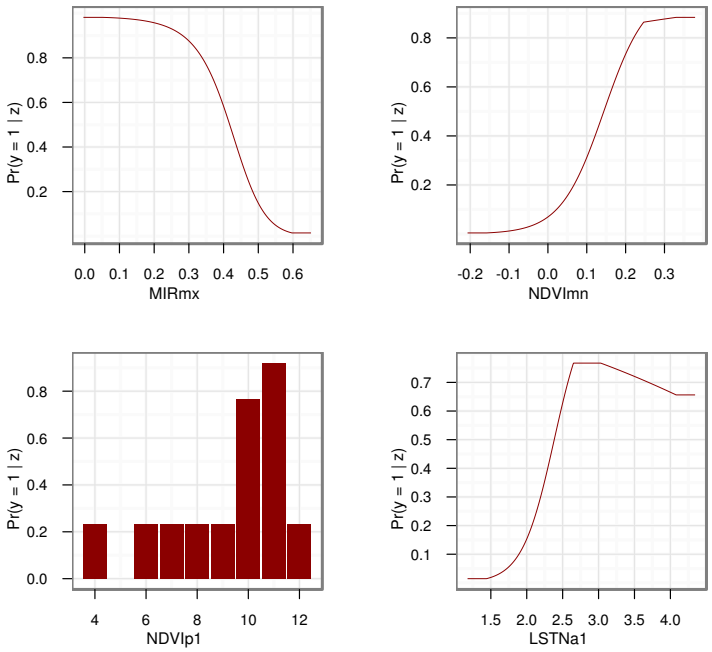
Graphique 4.6 : Importance des variables environnementales dans le modèle MaxEnt.

Le gain global est une mesure de qualité du modèle, associé au logarithme de la vraisemblance maximisée. Le gain mesuré avec le modèle complet (toutes les variables) est de 2.135 donc la distribution (calculée par MaxEnt) *fitte* le jeu de calibration 8 fois ($exp(2.135) \approx 8$) mieux qu'une distribution de probabilité uniforme. La variable qui contribue le plus globalement (au gain) est MIRmx(40%). Le MIRmx est alors la variable, qui utilisée toute seule, apporte le plus d'information (graphique 4.6). La seconde variable la plus importante est NDVIp1, il s'agit de la variable qui entraîne la plus grande baisse sur le gain lorsqu'elle n'est

2. méthode de reéchantillonage, on calcule une statistique à chaque tirage, en enlevant une observation de l'échantillon à chaque fois

pas utilisée. Il s'agit alors d'une variable qui contient des informations que n'apportent pas les autres variables.

Nous avons réalisé une analyse de l'effet marginal de chaque variable sur la probabilité d'occurrence. Toute chose égale par ailleurs, nous avons fait varier les quatre variables les plus importantes du modèles afin d'analyser la sensibilité du modèle et la réponse de l'espèce aux variations de chacune d'entre elles (graphique 4.7).



Graphique 4.7 : Effet marginal des variables sur la probabilité d'occurence.

Ceteris paribus, la probabilité de présence augmente lorsque le NDVIp1 tourne autour des mois d'octobre et novembre (mois 10 et 11) de chaque année. En effet, le NDVIp1 permet de discriminer la végétation arborée de la végétation herbacée qui ont des périodes de développement différents. Les glossines riveraines si elles ont pu se passer de leurs biotopes originels (galeries forestières), sont inféodées à une végétation arborée dense qui leur confère un ombre suffisante et tamponne la température ambiante. Toute choses égales par ailleurs, une augmentation du MIRmx correspond à une baisse de la probabilité de présence. On remarque aussi que les variables dont la courbe d'influence sur la probabilité est monotone (MIRmx, NDVImn) sont celles de marginalité alors que les courbes en cloches comme celle du LSTNa1 sont caractéristiques de la spécialisation, une partie de leur gamme de variation seulement est favorable à l'espèce.

Conclusions et perspectives

Au terme de ce document, on note que l'approche utilisée nous a permis de mettre en place un modèle conceptuel qui repose sur des bases théoriques solides.
Une analyse exploratoire des variables qui influencent la distribution de l'espèce a mis en évidence une partie des relations qui existent entre l'environnement et notre espèce d'étude, *G. p. gambiensis*. Cette analyse, a en outre permis de faire un choix de variables basé sur l'écologie de l'espèce. Nous avons trouvé que *G. p. gambiensis* est très sensible aux sols nus (sans végétation) en saison sèche, qu'elle doit éviter pour sa survie. La recherche de végétation arborée qui joue un rôle tampon pour la température à été aussi décelée par ces méthodes factorielles.
Ces résultats ont été confirmés et affinés par le modèle prédictif final, qui a permis en outre d'avoir une carte de distribution du vecteur de la TAA dans la zone des Niayes. Cette carte de distribution nous a permis d'actualiser la première zone de lutte établie avant cette étude.
La zone prédite par le meilleur modèle permet de tirer les conclusions suivantes :

- Les habitats favorables situés le long de la côte correspondent aux ravines d'évacuation de l'eau en saison des pluies. En particulier l'habitat favorable le plus important en contact avec le sud de la zone de lutte correspond à la rivière Somone qui relie cette dernière à la réserve de Bandia où les glossines étaient effectivement présentes dans les années 70 avant que cette population ne s'éteigne suite à une forte dégradation de la végétation arborée (Laveissière et Traoré, 1979). Cependant cette végétation a depuis été restaurée grâce à la mise en réserve de cette zone (réserve de Bandia).
- Certains habitats favorables qui ne sont pas en contact avec la zone de lutte, ne sont actuellement plus occupés par les glossines, notamment en raison d'une campagne d'éradication menée dans les années 70 (Touré, 1972). Il est possible que ces habitats, dénommés patches ou îlot dans le vocabulaire des métapopulations (Hanski et Gaggioti, 2004), soient trop éloignés de la zone infestée principale pour être colonisés par les glossines. On considère généralement que celles ci peuvent franchir une distance d'environ 2 km dans la matrice défavorable (Cuisance *et al.*, 1985 ; Bouyer *et al.*,

2010a). Cela renvoie au problème de l'accessibilité de l'habitat favorable : une des hypothèses écologiques fondamentales des modèles d'analyse de niche n'est donc pas respectée. Il serait dès lors intéressant de pouvoir intégrer la structure spatiale dans ce type de d'analyse.

– La qualité du modèle final pour la prédiction de l'habitat favorable est limitée par la résolution spatiale des données environnementales disponibles (1 km^2). En effet, les glossines peuvent parfois se rencontrer dans des microhabitats de quelques dizaines de mètres carré (Bouyer *et al.*, 2010b). Une des voies d'amélioration de ce modèle pourrait donc être l'intégration dans l'analyse de données environnementales à plus haute résolution spatiale. Ce travail a conduit le projet d'éradication à programmer l'utilisation d'images LandSat (30 m de résolution spatiale).
– La zone de lutte actuelle n'est globalement pas remise en cause ;
– Il est cependant nécessaire de réaliser des piégeages supplémentaires dans les habitats prédits en contact avec la zone de lutte au nord et au sud de celle ci.

Malgré quelques limitations liées aux données (résolution spatiale, nombre de points de présence) et aux modèles utilisés, cette étude a permis d'évaluer une distribution potentielle de *G. p. gambiensis*. Les cartes d'habitats favorables ainsi obtenues sont des outils décisionnels importants dans le cadre de la lutte intégrée contre les vecteurs de la TAA. Cependant, ces cartes sont statiques, et la niche des glossines riveraines est fortement influencée par les conditions climatiques. Ainsi parmi les perspectives de recherche, les point suivant semble intéressant et mérite des développements :

- La mise en place de modèle de distribution spatio-temporelle d'espèce
- Intégration de la théorie des métatpopulations (Hanski et Gaggioti, 2004) qui relâche beaucoup d'hypothèse restrictive des modèles classiques de distribution d'espèce

Ces développements permettraient alors aux programmes de luttes de disposer d'informations régulièrement mises à jour sur la distribution des vecteurs de TAA.

Bibliographie

Austin, M. (2002). Spatial prediction of species distribution : an interface between ecological theory and statistical modelling. *Ecological modelling*, 157(2-3) :101--118.

Austin, M. (2007). Species distribution models and ecological theory : a critical assessment and some possible new approaches. *Ecological modelling*, 200(1-2) :1--19.

Barclay, H. et Hargrove, J. (2005). Probability models to facilitate a declaration of pest-free status, with special reference to tsetse (diptera : Glossinidae). *Bulletin of entomological research*, 95(1) :1--12.

Basille, M., Calenge, C., Marboutin, É., Andersen, R. et Gaillard, J. (2008). Assessing habitat selection using multivariate statistics : some refinements of the ecological-niche factor analysis. *Ecological Modelling*, 211(1-2) :233--240.

Bloomfield, P. (2004). *Fourier Analysis of Time Series : An Introduction*. Wiley series in probability and statistics : Applied probability and statistics. John Wiley & Sons.

Bouyer, J. (2006). *Écologie et contrôle des glossines et épidémiologie des trypanosomoses animales africaines*. Thèse de doctorat, Université de Montpelllier II, FR.

Bouyer, J., Ravel, S., Guerrini, L., Dujardin, J., Sidibé, I., Vreysen, M., Solano, P. et De Meeûs, T. (2010a). Population structure of glossina palpalis gambiensis (diptera : Glossinidae) between river basins in burkina faso : Consequences for area-wide integrated pest management. *Infection, Genetics and Evolution*, 10(2) :321--328.

Bouyer, J., Seck, M., Sall, B., Ndiaye, E., Guerrini, L. et Vreysen, M. (2010b). Stratified entomological sampling in preparation for an area-wide integrated pest management program : the example of glossina palpalis gambiensis (diptera : Glossinidae) in the niayes of senegal. *Journal of medical entomology*, 47(4) :543--552.

Bouyer, J., Solano, P., De La Rocque, S., Desquesnes, M., Cuisance, D., Itard, J., Frézil, J. et Authié, E. (2010c). Trypanosomoses : control methods. *Infectious and parasitic diseases of livestock*.

Boyd, D. et Curran, P. (1998). Using remote sensing to reduce uncertainties in the global carbon budget : the potential of radiation acquired in middle infrared wavelengths. *Remote Sensing Reviews*, 16(4) :293--327.

Breiman, L. (1984). *Classification and regression trees*. The Wadsworth and Brooks-Cole statistics-probability series. Chapman & Hall.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2) :123--140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1) :5--32.

Brotons, L., Thuiller, W., Araújo, M. et Hirzel, A. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, 27(4) :437--448.

Browning, D., Beaupre, S. et Duncan, L. (2005). Using partitioned mahalanobis d2 (k) to formulate a gis-based model of timber rattlesnake hibernacula. *Journal of Wildlife Management*, 69(1) :33--44.

Calenge, C. (2006). The package adehabitat for the r software : tool for the analysis of space and habitat use by animals. *Ecological Modelling*, 197 :1035.

Calenge, C. et Basille, M. (2008). A general framework for the statistical exploration of the ecological niche. *Journal of Theoretical Biology*, 252(4) :674--685.

Calenge, C., Darmon, G., Basille, M., Loison, A. et Jullien, J. (2008). The factorial decomposition of the mahalanobis distances in habitat selection studies. *Ecology*, 89(2) : 555--566.

Cecchi, G., Mattioli, R., Slingenbergh, J. et De La Rocque, S. (2008). Land cover and tsetse fly distributions in sub-saharan africa. *Medical and Veterinary Entomology*, 22(4) :364--373.

Challier, A. et de Paris VI, U. (1973). *Ecologie de" Glossina palpalis gambiensis" Vanderplank, 1949 (Diptera, Muscidae) en Savane d'Afrique occidentale*. ORSTOM Paris, France :.

Chefaoui, R. et Lobo, J. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological modelling*, 210(4) :478--486.

Clark, J., Dunn, J. et Smith, K. (1993). A multivariate model of female black bear habitat use for a geographic information system. *The Journal of wildlife management*, pages 519--526.

Cohen, J. *et al.* (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1) :37--46.

Cressie, N. et Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Cuisance, D., Février, J., Dejardin, J. et Filledier, J. (1985). Dispersion linéaire de glossina palpalis gambiensis et de glossina tachinoides dans une galerie forestière en zone soudano-guinéenne (burkina-faso). *Revue d'Elevage et de Médecine vétérinaire des Pays tropicaux*, 38 :153--172.

Davison, A. et Hinkley, D. (1997). *Bootstrap methods and their application*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press.

DeLong, E., DeLong, D. et Clarke-Pearson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves : a nonparametric approach. *Biometrics*, pages 837--845.

Dunn, J. et Duncan, L. (2000). Partitioning mahalanobis d2 to sharpen gis classification. *In International conference on management information systems incorporating GIS & remote sensing*, pages 195--204.

Elith, J., Graham*, C., Anderson, R., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J., Lehmann, A. *et al.* (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2) :129--151.

Elith, J., Phillips, S., Hastie, T., Dudík, M., Chee, Y. et Yates, C. (2011). A statistical explanation of maxent for ecologists. *Diversity and Distributions*.

Elton, C. (1927). *Animal ecology*. University of Chicago Press.

Freeman, E. (2007). *PresenceAbsence : An R Package for Presence-Absence Model Evaluation*. USDA Forest Service, Rocky Mountain Research Station, 507 25th street, Ogden, UT, USA. eafreeman@fs.fed.us.

Gabriel, K. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3) :453--467.

Genuer, R. (2010). *Forêts aléatoires : aspects théoriques, sélection de variables et applications*. Thèse de doctorat, Université PARIS-SUD XI, FR.

Grinnell, J. (1917). The niche-relationships of the california thrasher. *The Auk*, 34(4) :427--433.

Guerrini, L. (2009). *Le risque trypanosomien dans le bassin du Mouhoun au Burkina Faso : approches paysagères*. Thèse de doctorat, Université de Montpelllier III, FR.

Guis, H. (2007). *Géomatique et épidémiologie : caractérisation des paysages favorables à Culicoides imicola, vecteur de la fièvre catarrhale ovine en Corse*. Thèse de doctorat, Université de Franche-Comté.

Guisan, A. et Thuiller, W. (2005). Predicting species distribution : offering more than simple habitat models. *Ecology letters*, 8(9) :993--1009.

Guisan, A. et Zimmermann, N. (2000). Predictive habitat distribution models in ecology. *Ecological modelling*, 135(2-3) :147--186.

Hall, L., Krausman, P. et Morrison, M. (1997). The habitat concept and a plea for standard terminology. *Wildlife Society Bulletin*, pages 173--182.

Hanski, I. et Gaggioti, . E. (2004). *Ecology, genetics and evolution of metapopulations*. Elsevier Academic Press.

Hargrove, J. (1988). Tsetse : the limits to population growth. *Medical and veterinary Entomology*, 2(3) :203--217.

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer.

Hay, S. *et al.* (1997). Remote sensing and disease control : past, present and future. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 91(2) :105.

Hay, S., Randolph, S. et Rogers, D. (2000). *Remote sensing and geographical information systems in epidemiology*, volume 47. Academic Pr.

Hay, S., Tucker, C., Rogers, D., Packer, M. *et al.* (1996). Remotely sensed surrogates of meteorological data for the study of the distribution and abundance of arthropod vectors of disease. *Annals of Tropical Medicine and Parasitology*, 90(1) :1--20.

Hendrickx, G., Napala, A., Dao, B., Batawui, D., De Deken, R., Vermeilen, A. et Slingenbergh, J. (1999a). A systematic approach to area-wide tsetse distribution and abundance maps. *BULLETIN OF ENTOMOLOGICAL RESEARCH-LONDON-*, 89 :231--244.

Hendrickx, G., Napala, A., Dao, B., Batawui, K., Bastiaensen, P., De Deken, R., Vermeilen, A., Vercruysse, J. et Slingenbergh, J. (1999b). The area-wide epidemiology of bovine trypanosomosis and its impact on mixed farming in subhumid west africa ; a case study in togo. *Veterinary parasitology*, 84(1-2) :13--31.

Hijmans, R. J., Phillips, S., Leathwick, J. et Elith, J. (2011). *dismo : Species distribution modeling*. R package version 0.7-8.

Hirzel, A., Hausser, J., Chessel, D. et Perrin, N. (2002). Ecological-niche factor analysis : how to compute habitat-suitability maps without absence data ? *Ecology*, 83(7) :2027--2036.

Hirzel, A. et Le Lay, G. (2008). Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, 45(5) :1372--1381.

Hirzel, A., Le Lay, G., Helfer, V., Randin, C. et Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, 199(2) :142--152.

Hutchinson, G. (1957). Concluding remarks. *In Cold Spring Harbor Symposia on Quantitative Biology*, volume 22, pages 415--427. Cold Spring Harbor Laboratory Press.

Hutchinson, G. (1978). *An introduction to population ecology*, volume 260. Yale University Press New Haven, Connecticut, USA.

Itard, J., Cuisance, D. et Tacher, G. (2003). Trypanosomoses : historique--répartition géographique. *Principales maladies infectieuses et parasitaires du bétail. Europe et Régions chaudes, Lavoisier, Londres-Paris-New York, Éditions Tec et Doc/Éditions médicales internationales*, pages 1607--1615.

Jackson, C. (1941). The economy of a tsetse population. *Bulletin of entomological research*, 32(01) :53--55.

Knick, S. et Dyer, D. (1997). Distribution of black-tailed jackrabbit habitat determined by gis in southwestern idaho. *The Journal of wildlife management*, pages 75--85.

Laveissière, C. et Traoré, T. (1979). Enquête entomologique dans le foyer de trypanosomiase humaine de la somone (république du sénégal - mai 1979).

Lebart, L., Morineau, A. et Piron, M. (2008). *Analyse exploratoire multidimensionnelle*. Dunod.

Liu, C., Berry, P., Dawson, T., Pearson, R. *et al.* (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28(3) :385--393.

Mahalanobis, P. (1936). On the generalized distance in statistics. *In Proceedings of the National Institute of Science, Calcutta*, volume 12, page 49.

Meynard, C. et Quinn, J. (2007). Predicting species distributions : a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, 34(8) :1455--1469.

Morrison, M., Marcot, B. et Mannan, R. (2006). *Wildlife-habitat relationships : concepts and applications*. Island Pr.

Nash, T. (1937). Climate, the vital factor in the ecology of glossina. *Bulletin of entomological research*, 28(01) :75--127.

Nash, T. (1948). Tsetse flies in british west africa : the gold coast. *Her Majesty Stationary Office, London, UK*, 28 :47.

Neteler, M. (2010). *Spatio-temporal reconstruction of satellite-based temperature maps and their application to the prediction of tick and mosquito disease vector distribution in Northern Italy*. BoD--Books on Demand.

Neteler, M. et Mitasova, H. (2008). *Open source GIS : a GRASS GIS approach*. Kluwer international series in engineering and computer science. Springer.

Perpiñán Lamigueiro, O. et Hijmans, R. (2011). *rasterVis : Visualization methods for the raster package*. R package version 0.10-7.

Phillips, S., Anderson, R. et Schapire, R. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4) :231--259.

R Development Core Team (2011). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reddy, M. (2006). *Textbook of remote sensing and geographical information systems*. BS Publications.

Robinson, T., Rogers, A. et Williams, B. (1997a). Mapping tsetse habitat suitability in the common fly belt of southern africa using multivariate analysis of climate and remotely sensed vegetation data. *Medical and Veterinary Entomology*, 11 :235--245.

Robinson, T., Rogers, A. et Williams, B. (1997b). Univariate analysis of tsetse habitat in the common fly belt of southern africa using climate and remotely sensed vegetation data. *Medical and Veterinary Entomology*, 11 :223--234.

Rogers, D. (1979). Tsetse population dynamics and distribution : a new analytical approach. *The Journal of Animal Ecology*, pages 825--849.

Rogers, D. (2000). Satellites, space, time and the african trypanosomiases. *Advances in Parasitology*, 47 :129--171.

Rogers, D., Hay, S. et Packer, M. (1996). Predicting the distribution of tsetse flies in west africa using temporal fourier processed meteorological satellite data. *Annals of Tropical Medicine and Parasitology*, 90(3) :225--242.

Rogers, D. et Randolph, S. (1991). Mortality rates and population density of tsetse flies correlated with satellite imagery. *Nature*.

Rogers, D. et Randolph, S. (1993). Distribution of tsetse and ticks in africa : past, present and future. *Parasitology Today*, 9(7) :266--271.

Rotenberry, J., Knick, S. et Dunn, J. (2002). A minimalist approach to mapping species' habitat.

Rotenberry, J., Preston, K. et Knick, S. (2006). Gis-based niche modeling for mapping species'habitat. *Ecology*, 87(6) :1458--1464.

Rudin, W. (1987). *Real and complex analysis*. Mathematics series. McGraw-Hill.

Scharlemann, J., Benz, D., Hay, S., Purse, B., Tatem, A., Wint, G. et Rogers, D. (2008). Global data for ecology and epidemiology : a novel algorithm for temporal fourier processing modis data. *PLoS One*, 3(1) :e1408.

Seck, M., Bouyer, J., Sall, B., Bengaly, Z. et Vreysen, M. (2010). The prevalence of african animal trypanosomoses and tsetse presence in western senegal. *Parasite*, 17(3) : 257--265.

Soberón, J. (2007). Grinnellian and eltonian niches and geographic distributions of species. *Ecology Letters*, 10(12) :1115--1123.

Soberón, J. et Peterson, A. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity informatics*, 2(0).

Touré, S. (1974). Notes sur quelques particularités dans l'habitat de glossina palpalis gambiensis vanderplank 1949 (diptera : Glossinidae) observées au sénégal. *Rev Elev Méd vét Pavs trop*, 27 :81--91.

Touré, S. (1972). Lutte contre glossina palpalis gambiensis dans la région des niayes du sénégal. *Revue d'Elevage et de Médecine vétérinaire des Pays tropicaux*, 22 :339--347.

Tran, A. (2004). *Télédétection et Épidémiologie : Modélisation de la dynamique de populations d'insectes et application au contrôle de maladies à transmission vectorielle [Thèse de Doctorat, thesis]*. Thèse de doctorat, Université Louis Pasteur Strasbourg.

Tran, A., Biteau-Coroller, F., Guis, H. et Roger, F. (2005). Modélisation des maladies vectorielles. *Epidémiol et Santé Anim*, 47 :35--51.

van Etten, R. J. H. . J. (2011). *raster : Geographic analysis and modeling with raster data*. R package version 1.9-55.

Vancutsem, C., Ceccato, P., Dinku, T. et Connor, S. (2010). Evaluation of modis land surface temperature data to estimate air temperature in different ecosystems over africa. *Remote Sensing of Environment*, 114(2) :449--465.

Wan, Z. (1999). Modis land-surface temperature algorithm theoretical basis document (lst atbd). *Institute for Computational Earth System Science, Santa Barbara*, page 75.

Wan, Z. (2006). Modis land surface temperature products users' guide. *Institute for Computational Earth System Science (ICESS), University of California, Santa Barbara*.

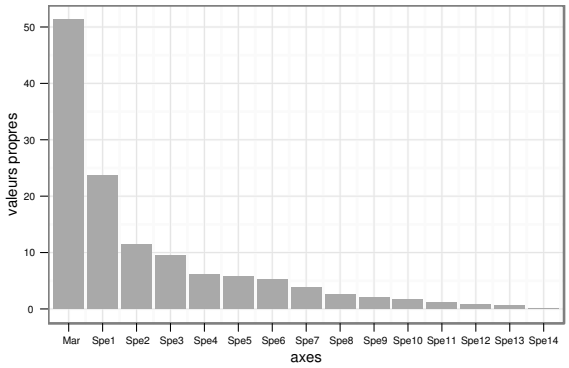
Ward, G., Hastie, T., Barry, S., Elith, J. et Leathwick, J. (2009). Presence-only data and the em algorithm. *Biometrics*, 65(2) :554--563.

Whittaker, R., Levin, S. et Root, R. (1973). Niche, habitat, and ecotope. *The American Naturalist*, 107(955) :321--338.

Wickham, H. (2009). *ggplot2 : elegant graphics for data analysis*. Springer New York.

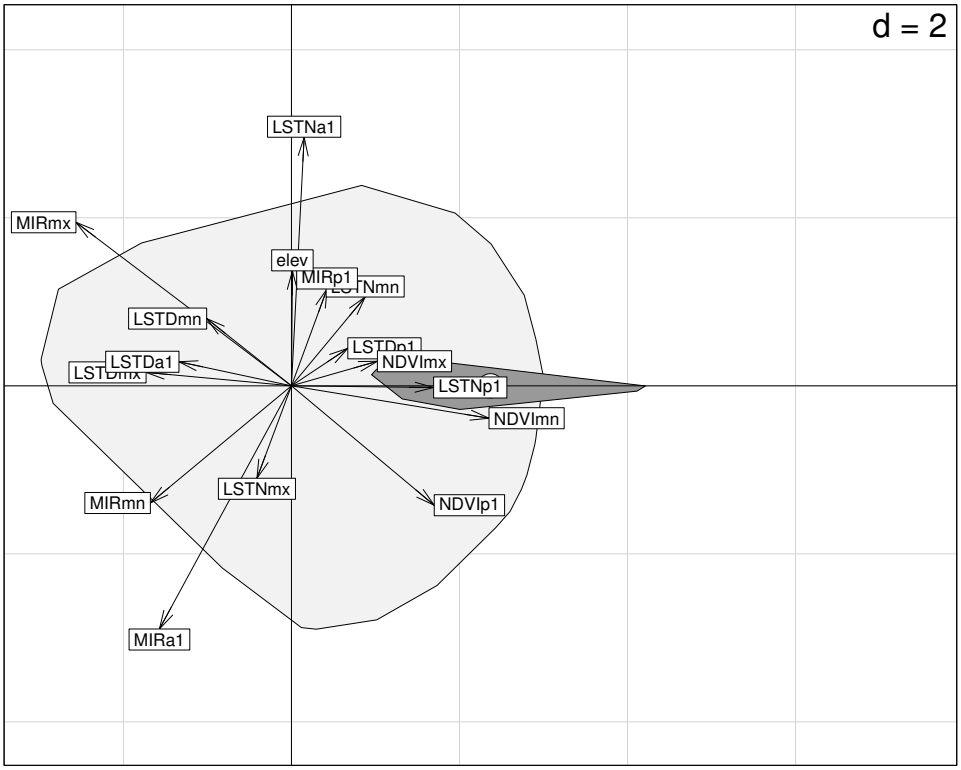
Annexe A

Valeur propre de l'ENFA :



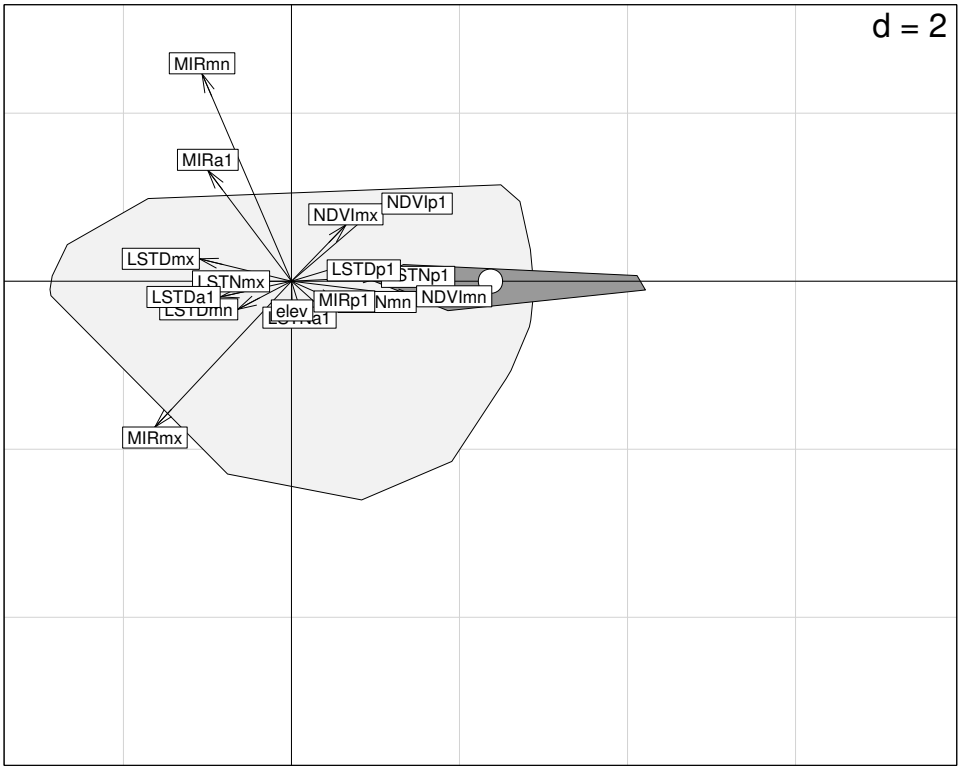
Graphique 1 : Valeurs propres de l'ENFA

Premier plan factoriel de l'ENFA :



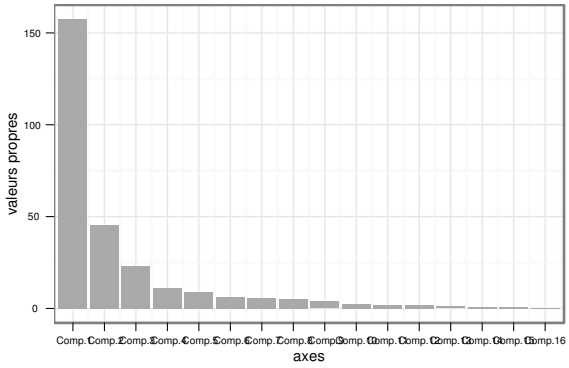
Graphique 2 : Biplot du premier plan factoriel, visualisation à l'aide l'enveloppe convexe minimale

Second plan factoriel de l'ENFA :



Graphique 3 : Biplot du second plan factoriel

Valeurs propres de la MADIFA :



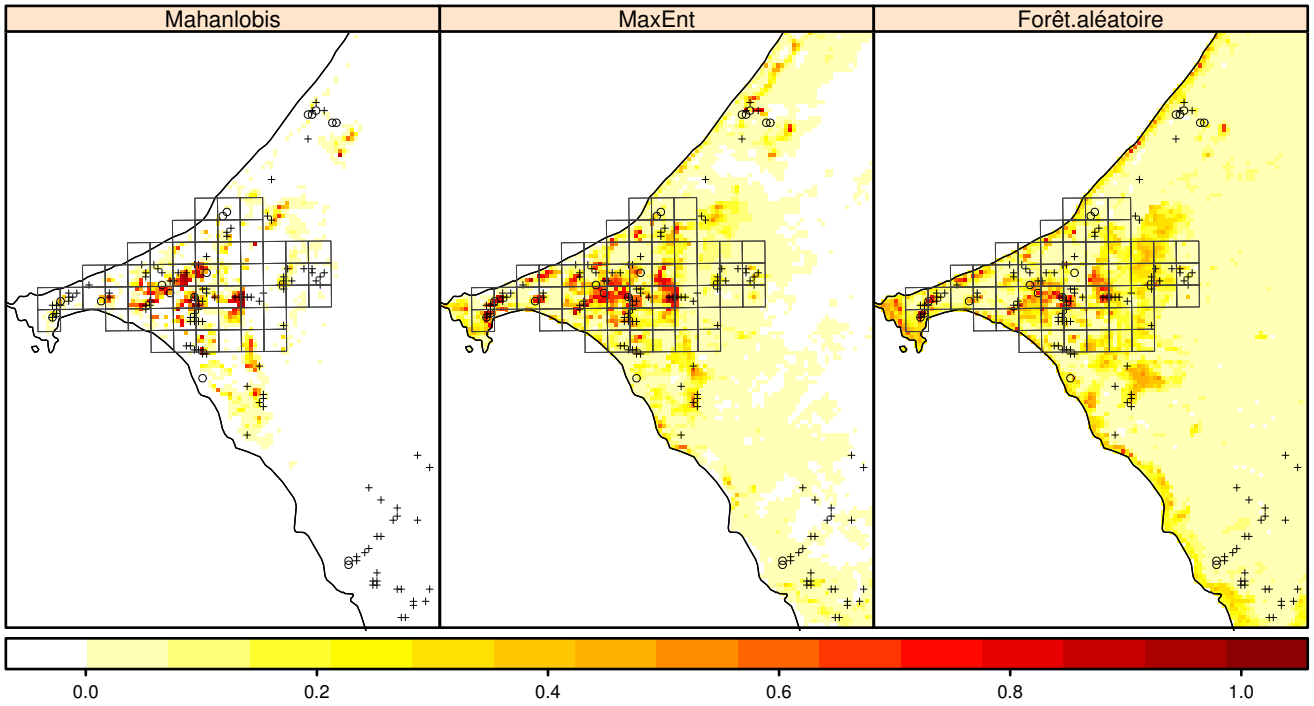
Graphique 4 : Valeurs propres de la MADIFA

Annexe B

Nous disposons d'une matrice de confusion

	Observé	
prédit	positif	négatif
positif	a	b
négatif	c	d

1. spécificité = $\frac{d}{b+d}$
2. sensibilité = $\frac{a}{a+c}$



Graphique 5 : Probabilité d'occurence de *G. p. gambiensis* avec relevés phytosociologiques. o pour gîtes favorables et + pour gîtes défavorables

Annexe C : logiciels utilisés

Cette étude a été possible grâce à l'utilisation de plusieurs logiciels.

Le logiciel principal utilisé est *R* (R Development Core Team, 2011) mais pour stocker les images rasters et faire certaines analyses nous avons utilisé le SIG GRASS (Neteler et Mitasova, 2008).

Les différents packages de *R* utilisés sont :

- **raster** (van Etten, 2011) pour l'analyse des images de type raster.
- **dismo** (Hijmans *et al.*, 2011) pour certains modèles de distribution d'espèce.
- **adehabitatHS** (Calenge, 2006) a permi de réaliser les différentes analyses factorielles.
- **PresenceAbsence** (Freeman, 2007) pour la mesurer la qualité de différents modèles.
- **spgrass6** (Neteler et Mitasova, 2008) et **GrassScriptHelper** (crée par l'auteur, pas en ligne) pour avoir une interface complète entre *R* et GRASS.
- **ggplot2** (Wickham, 2009), **rasterVis** (Perpiñán Lamigueiro et Hijmans, 2011) ont permis de réaliser les différents graphiques.

Le texte a été écrit en utilisant XeLaTeX.